Cross-Lingual and Low-Resource Sentiment Analysis

Noura Farra

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Cross-Lingual and Low-Resource Sentiment Analysis

Noura Farra

Identifying sentiment in a low-resource language is essential for understanding opinions internationally and for responding to the urgent needs of locals affected by disaster incidents in different world regions. While tools and resources for recognizing sentiment in high-resource languages are plentiful, determining the most effective methods for achieving this task in a low-resource language which lacks annotated data is still an open research question. Most existing approaches for cross-lingual sentiment analysis to date have relied on high-resource machine translation systems, large amounts of parallel data, or resources only available for Indo-European languages.

This work presents methods, resources, and strategies for identifying sentiment cross-lingually in a low-resource language. We introduce a cross-lingual sentiment model which can be trained on a high-resource language and applied directly to a low-resource language. The model offers the feature of lexicalizing the training data using a bilingual dictionary, but can perform well without any translation into the target language.

Through an extensive experimental analysis, evaluated on 17 target languages, we show that the model performs well with bilingual word vectors pre-trained on an appropriate translation corpus. We compare in-genre and in-domain parallel corpora, out-of-domain parallel corpora, in-domain comparable corpora, and monolingual corpora, and show that a relatively small, in-domain parallel corpus works best as a transfer medium if it is available. We describe the conditions under which other resources and embedding generation methods are successful, and these include our strategies for leveraging in-domain comparable corpora for cross-lingual sentiment analysis.

To enhance the ability of the cross-lingual model to identify sentiment in the target language, we present new feature representations for sentiment analysis that are incorporated in the cross-lingual model: bilingual sentiment embeddings that are used to create bilingual sentiment scores, and a method for updating the sentiment embeddings during training by lexicalization of the target language. This feature configuration works best for the largest number of target languages in both untargeted and targeted cross-lingual sentiment experiments.

The cross-lingual model is studied further by evaluating the role of the source language, which has traditionally been assumed to be English. We build cross-lingual models using 15 source languages, including two non-European and non-Indo-European source languages: Arabic and Chinese. We show that language families play an important role in the performance of the model, as does the morphological complexity of the source language.

In the last part of the work, we focus on sentiment analysis towards targets. We study Arabic as a representative morphologically complex language and develop models and morphological representation features for identifying entity targets and sentiment expressed towards them in Arabic open-domain text. Finally, we adapt our cross-lingual sentiment models for the detection of sentiment towards targets. Through cross-lingual experiments on Arabic and English, we demonstrate that our findings regarding resources, features, and language also hold true for the transfer of targeted sentiment.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deep gratitude to my advisor, Kathleen McKeown, for her guidance, support, and patience, and for creating an environment where I felt I could express my ideas and pursue the research direction I was most passionate about. I have learned so much from Kathy and this thesis would not have been possible without her. I also wish to thank the rest of my thesis committee members, Smaranda Muresan, Mona Diab, Julia Hirschberg, and Owen Rambow, and Michael Collins, who was on my proposal committee, for their advice and support.

I am thankful to Nizar Habash, who advised me during my first two years of my PhD and who kept my funding for two years after he left. Nizar accepted me into Columbia University, introduced me to Arabic natural language processing, and taught me a lot about doing research. This thesis also would not have been possible without him.

Throughout the last seven years, I have been fortunate enough to meet and work with dear friends and colleagues who have made my time at Columbia and New York a memorable one, and who have not hesitated to provide help or advice when I needed it. I especially wish to thank Victor Soto, Ghada Almashaqbeh, Chris Kedzie, Mohammad Sadegh Rasooli, Jessica Ouyang, Chris Hidey, Elsbeth Turcan, Fei-Tzin Lee, Emily Allaway, Gaurav Gite, Hooshmand Shokri, Ramy Eskander, Tuhin Chakrabarty, Tariq Alhindi, Tom Effland, Olivia Winn, Andrea Lottarini, Sara Rosenthal, Or Biran, Kapil Thadani, Yves Petinot, Brenda Yang, Shabnam Tafreshi, and Efsun Kayi. For those early days and unforgettable times we spent at CCLS, I thank Boyi Xie, Heba Alfardy, Wael Salloum, Vinod Prabhakaran, Apoorv Agarwal, Weiwei Guo, John Min, Swabha Swayamdipta, Nadi Tomeh, Mohamed Altantawy, and Ahmed Kholy.

I am sincerely grateful to the family and friends who made my stay in New York feel like home. A most special thanks to Wassim and Dahlia, who have been there for me through the good times and the bad, for Thanksgiving dinners, and for giving me a home away from home. My dear thanks to my great-uncle and aunt, to Toufic, and to Sonali, all of whom have made New York so much more enjoyable, and to my best friends in Beirut, Nadine and Chirine, from whom distance has not kept me apart.

To the two people who have been behind me every step of the way: words cannot express the extent of my love and gratitude to my mother and father. I know that my PhD has not been easy for them either, and that they have lived my struggles as though they were their own. I would not be where I am today without them. Of course, no thesis acknowledgment would be complete without mentioning Akram and Rouba, who are definitely in my list of top favorite couples, and my nephew and niece, Jad and Judy, who have brought joy to my life in the last seven years. Jado is only a year older than my PhD, but that doesn't mean he won't be reading these acknowledgments.

Finally, the long journey through a doctorate degree is a creative and independent one, but it can also be an lonely one. To the Graduate Workers of Columbia, my deepest thanks. Since our union election in 2016, GWC has given me a voice and made me feel part of something bigger than myself. Because of GWC, I've made friends throughout schools across the university, and they've helped me (among other things!) finally learn the names and locations of campus buildings outside the engineering school. With their courage, spirit, and camraderie, they have reminded me that I am not alone, and that we are stronger when we're united than when we're divided. I'm honored to have been part of this fight.

To those unheard, who continue to express their sentiments.

Chapter 1

Introduction

"Real names tell you the story of the things they belong to in my language, in the Old Entish as you might say. It is a lovely language, but it takes a very long time to say anything in it, because we do not say anything in it, unless it is worth taking a long time to say, and to listen to."

— J. R. R. Tolkien, *The Lord of the Rings: One Volume*

There are more than seven thousand known living languages in the world. Yet, the number of languages that has been studied in terms of computational linguistics is probably fewer than thirty, as the vast majority of known living languages lack the computational and linguistic annotation resources required for building natural language processing systems (Maxwell and Hughes, 2006; Baumann and Pierrehumbert, 2014). Indeed, about half of the world speaks a language not in the top twenty most commonly spoken languages (Lewis, 2009; Littell et al., 2018), and even these most spoken languages are not all equipped with the rich resources required for building complex machine learning models that can recognize and identify human sentiment.

The ability of a computionally-driven system to identify sentiment in a new language, however, is necessary if we are to build machine learning systems that can aggregate and understand human opinions from all parts of the world. The dominantly spoken language in a given region of the world, and the linguistic and computational resources available for it, is often determined by political factors, such as the governing power or the dominant ethnic group. The Uyghurs, for example, are a Muslim minority in China who live in the Xinjiang Autonomous Region and speak the Uyghur language, a Turkic language with about 10 to 15 million speakers, among them several ethnic minorities in Xinjiang. The Tigriyans, who speak the low-resource language Tigrinya, are an ethnic minority in the Tigray region of Ethiopia, where the dominant spoken language is Amharic. In the Arab world, Modern Standard Arabic is used as the official language by all Arab countries, while the true languages spoken in practice by everyday locals are the dialects, which are themselves in essence low-resource languages, as they are spoken far more often than they are written.

However, with the fast-growing rise of social media platforms such as Twitter and Facebook, and their use by millions of people around the world to relate their personal and affectual experiences, it is precisely the languages spoken by the locals which matter most when it comes to expressing sentiment. This is especially the case when a natural disaster or a significant political incident occurs in one part of the world, such as the earthquake that struck the Xinjiang region in August 2017, or the ethnic conflict that occurred in Ethiopia between 2015 and 2017, and it is desired to assess the needs of locals in the most affected areas, or to accurately represent the views and reactions of residents to political events.

With the majority of resources and studies dedicated to sentiment analysis still currently concentrated towards a few high-resource languages, the task of identifying sentiment in a new, poorly or even moderately resourced language remains a challenge. Traditionally, the problem of assembling sentiment models for languages other than English has been approached using machine translation, (e.g. Balahur and Turchi (2014); Zhou et al. (2016)): translating datasets and corpora from or into the new target language and making use of high-performing sentiment analysis models that have already been developed for English. The machine translation solution falls short, however, when considering languages which lack resources to build such complex systems - machine translation models operating on neural network architectures require large amounts of manually created human translations, often in the order of millions.

In this work, we approach the problem of low-resource sentiment analysis from a *cross-lingual* perspective: using labeled datasets and text corpora from a more highly resourced *source* language to transfer sentiment to a low-resource *target* language, that lacks a labeled dataset. In this cross-lingual sentiment approach, we are joined by more recent studies (e.g Zhou et al. (2014); Chen et al. (2016)). However, our work is distinct in several aspects. First, our models make use of untraditional resources for bridging source and target languages; these include comparable corpora that are not necessarily composed of direct translations, parallel corpora obtained from religious texts in the source and target languages, and relatively small sizes of other translation corpora. Second, we develop and provide an analysis of cross-lingual sentiment models using *source* languages which are themselves more poorly resourced than English, such as Arabic, Chinese, and a number of moderately-resourced European languages. Third, we employ various new techniques and strategies for generating cross-lingual feature representations for our models, including lexicalization of the target language and cross-lingual pre-training of sentiment features, and we extensively compare different cross-lingual feature representations for the task depending on the nature and availability of resources. Finally, we also study *targeted* cross-lingual sentiment analysis, where the cross-lingual model predicts not just the overall the sentiment of the text but also the sentiment towards a given topic, or target, a problem which remains unexplored by most previous work.

1.1 Contributions of the Work

The thesis presents both techniques as well as extensive experimental analyses towards achieving effective cross-lingual sentiment analysis, with a focus on the following factors: the nature of resources and their availability, efficient representation of bilingual features, the role and specificities of the source language (e.g, morphological complexity), and the application towards sentiment targets. As such we make the following contributions:

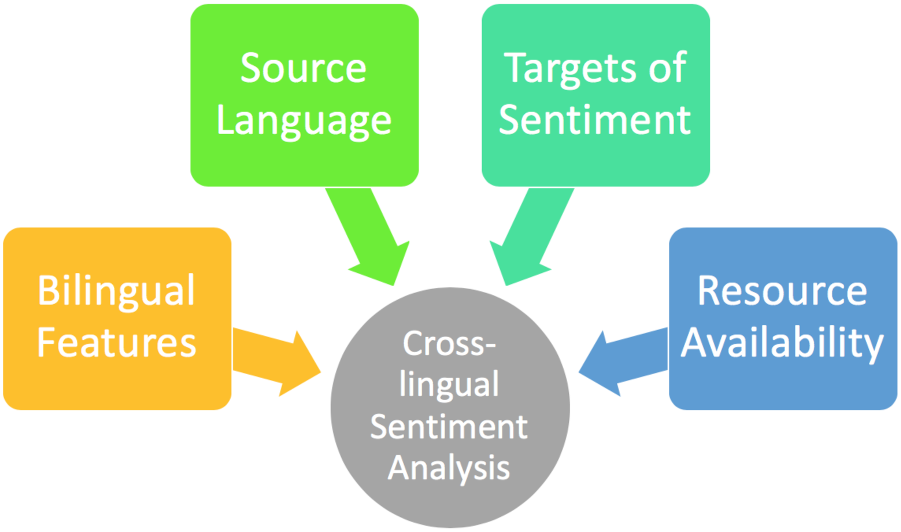


Figure 1.1: Overview of thesis topics.

1. A cross-lingual sentiment transfer model, trained on a high-resource or moderately-resourced source language, and applied to a low-resource target language. The model offers the feature of lexicalizing the training data using a bilingual dictionary, but can perform well without any translation into the target language.
2. The effective use of untraditional resources, including non-parallel comparable corpora, for training the cross-lingual model.
3. A detailed experimental analysis, evaluated on 17 target languages, of the cross-lingual word representation features and embedding generation methods best suited for the cross-lingual sentiment task. As part of this analysis, we compare the performance of word embedding vectors generated from: (a) In-genre and

in-domain parallel corpora, (b) Out-of-domain parallel corpora, (c) In-domain comparable corpora, and (d) Monolingual corpora with and without access to a bilingual dictionary. We show that pre-training bilingual features directly on a relatively small, in-domain parallel corpus works best if it is available, and we present recommendations for alternatives when it is not, describing the conditions under which other resources are successful. Also included in this analysis is a new method we present for pre-training bilingual sentiment embeddings on a translation corpus and updating them during training using target language lexicalization. This method relies only on a source-language sentiment lexicon and is especially helpful for identifying sentiment in the target language when the occurrence of sentiment in the source language is different than that of the target language; for example, when the target evaluation data is skewed towards negative sentiment.

4. An experimental analysis, evaluated on 17 target languages, of the role of the source language when transferring sentiment cross-lingually; in this we study the best suited language pairs for cross-lingual sentiment among Indo-European language families, and we compare the performance of English, Arabic, and Chinese source languages in transferring to other target languages while controlling for resource availability. We find that that language families play an important role in the performance of cross-lingual sentiment models, as does the morphological complexity of the source language and the specificities of its training data.
5. A study of targeted sentiment analysis in Arabic as a representative morphologically complex language, and the role of morphological preprocessing and segmentation techniques in the identification of entity targets in Arabic and the sentiment expressed towards them.
6. An adaptation of cross-lingual sentiment models for the transfer of targeted

sentiment, where targeted cross-lingual sentiment models are trained on English and Arabic making use of the methods described in the work.

As part of the thesis, we also make available a number of resources and datasets:

1. A comparable corpus of Wikipedia articles collected for 18 languages queried for 61 broad topics and named entities, aligned on the topic level and the document level for language-linked articles.
2. Four native annotated sentiment evaluation datasets for Chinese, Tigrinya, Uyghur, and Sinhalese.
3. Three new sentiment analysis training and evaluation datasets for Arabic: untargeted and targeted sentiment analysis datasets collected as part of our work organizing SemEval 2017 Task 4: Sentiment Analysis in Twitter (Rosenthal et al., 2017), and a targeted dataset annotated for sentiment towards entities in online comments to Arabic newspaper articles (Farra et al., 2015a).

1.2 Organization of the Thesis

The thesis is composed of eight chapters including the introductory chapter. We review related work in the field in Chapter 2. The progression of subsequent chapters is aimed to reflect the steps carried out in the process of building an end-to-end cross-lingual sentiment analysis application for a target language. Chapter 3 describes the collection of cross-lingual sentiment resources used throughout the work, which vary in availability depending on the target language considered. These resources include datasets for targeted and untargeted sentiment analysis, as well as the bilingual and monolingual resources, several of them unconventional, that we use to bridge the gap between source and target languages during transfer.

Chapter 4 presents our models for cross-lingual sentiment transfer using English as a source language, along with our experimental analysis of bilingual features best suited for the task (Farra and McKeown, 2019; Rasooli et al., 2018). Chapter 5 studies the role of the source language by presenting our work on transferring sentiment from languages other than English, along with a focus on the role of preprocessing when using a morphologically rich source language such as Arabic.

We turn to targeted sentiment analysis in Chapter 6, where we describe our collection of an Arabic targeted sentiment dataset of short documents, and our approach for identifying target entities and the sentiment towards them in Arabic, bearing in mind the segmentation techniques that work best for the language (Farra et al., 2015a; Farra and McKeown, 2017). Finally, Chapter 7 builds on the whole work with the goal of assembling cross-lingual systems for targeted sentiment analysis, trained on English and Arabic, and the thesis concludes in Chapter 8.

Before proceeding to next chapters, we present an overview of background and terminology related to the topic in Section 1.3.

1.3 Background and Terminology

We present a brief overview of background and terminology related to the task of sentiment analysis, and our classification of low-resource and high-resource languages that will be assumed throughout the work.

1.3.1 Sentiment Analysis

The task of sentiment analysis has been used interchangeably with opinion analysis, subjectivity analysis, and related tasks, e.g emotion analysis. When applied to text, a sentiment analysis system predicts the sentiment expressed in the text, usually by the writer or another entity mentioned in the text. In this work, we refer to the UNTARGETED sentiment task as the task of predicting the general or overall sentiment expressed by the text, while the TARGETED sentiment task is to predict the sentiment

Example 1. Thousands of refugees were turned back at the border.
Example 2. Refugees are facing many difficulties :(
Example 3. Refugees are taking over our jobs!!

Table 1.1: Examples of input text for sentiment analysis.

expressed specifically towards a given topic, such as an entity, or a *situation*, such as the need for urgent rescue during an earthquake.

Sentiment may be expressed on a number of scales; this work considers untargeted and targeted sentiment on three-point and two-point scales: 'positive', 'negative', and 'neutral', or 'positive' and 'negative'. Some other work (Wilson et al., 2005) has additionally separated the categories of 'neutral' and 'subjective-neutral'; in our work, since we are interested only in discerning positive and negative sentiment from all other categories, we consider all text without positive or negative sentiment to fall in the 'neutral' category.

In Table 1.1, Example 1 expresses neutral sentiment, Example 2 expresses negative untargeted sentiment, and Example 3 expresses negative targeted sentiment towards the entity *refugees*.

1.3.2 Low, Moderate, and High-Resource Languages

There have been different definitions regarding what constitutes a 'low-resource' language. The term 'low-density' has been used to describe languages for which very few NLP computational resources and linguistic annotation resources (Maxwell and Hughes, 2006; Hogan, 1999) or online resources (Megerdoomian and Parvaz, 2008) exist. Maxwell and Hughes (2006) list a number of types of linguistically annotated resources which are scarce for low-density languages, such as availability of parallel text, text annotated for named entities, morphologically analyzed text, text marked for word boundaries and part of speech tags, syntactic and treebank annotations, semantically tagged text (e.g FrameNet), and dictionaries and lexical resources.

Language	**Classification**	**Wikipedia**	**Google Translate**
English, Spanish, German	High	1M+	yes
Arabic, Persian, Russian	Moderate	100k+	yes
Ugyhur	Low	1k+	no
Tigrinya	Low	0.1k+	no

Table 1.2: Examples of high-resource, moderately-resourced, and low-resource languages, with approximate number of available Wikipedia articles, and availability of Google Translate.

For our sentiment task, however, we require that another resource be present: availability of a labeled *training dataset*. This is required to build supervised sentiment analysis systems using state-of-the-art natural language processing models, which require at least thousands of annotated data samples. In fact, when it comes to TARGETED sentiment analysis, which requires a more fine-grained annotation dataset, very few languages satisfy this requirement, and the importance of methods for cross-lingual sentiment analysis is then even more significant.

Cieri et al. (2016) makes the distinction between low-density languages and 'critical' languages, where the supply of resources does not meet demands. The Critical Language Scholarship program for 2015 listed several such languages, among them Arabic, Chinese, and Russian. In this work we refer to such languages, which have smaller amounts of sentiment training data or parallel translation corpora, as 'moderately-resourced' languages with respect to natural language processing and sentiment analysis resources. The morphological complexity of some of these languages (e.g Arabic, Russian) means that additional effort is required to develop language processing and sentiment analysis pipelines.

Throughout this work, therefore, low-resource languages will refer to languages that have no or very few linguistically annotated resources available for sentiment analysis, namely sentiment training datasets and parallel text used for bridging the language gap. Moderately-resourced languages have smaller amounts of such resources, but are also used as source languages for their potential of transferring to target languages with similar properties. Table 1.2 shows examples of our classifications.

Chapter 2

Related Work

"Progress lies not in enhancing what is, but in advancing toward what will be."

— Gibran Khalil Gibran, *A Handful of Sand on the Shore*

Sentiment analysis in text has been one of the fastest growing areas of study and application in natural language processing in the last twenty years, with applications in product and customer review analysis (Hu and Liu, 2004; Pontiki et al., 2014), social media and Twitter analysis (Agarwal et al., 2011; Pak and Paroubek, 2010), and other related tasks in language processing as well as in financial, political and social sciences. Many resources have been developed accordingly for the analysis of sentiment in English, including training datasets (Rosenthal et al., 2015b), lexicons annotated for sentiment and subjectivity (Wilson et al., 2005), lexical and semantic knowledge bases (Miller, 1995; Baccianella et al., 2010), as well as phrase-level sentiment annotations (Socher et al., 2013). The work described in this survey is focused on *cross-lingual* methods that have attempted to make use of the rich resources in English to build natural language processing systems that can identify sentiment in a low-resource language. We describe past work as it relates to cross-lingual methods, cross-lingual representations used, and resources. In addition, since our work addresses *targets* of sentiment, we describe past work in high-resource and cross-lingual targeted sentiment analysis, including the creation of resources for annotating targeted sentiment.

We start our survey with cross-lingual sentiment analysis (Section 2.1), describing

Tasks	Setting	Method	Notable Work
Untargeted	Cross-lingual	Machine Translation	**Co-train** (Wan, 2009) **English-to-target** (Balahur and Turchi, 2014) **Target-to-English** (Salameh et al., 2015)
		Annotation Projection	**Parallel Projection** (Mihalcea et al., 2007)
		Mixture Models	**CLMM** (Meng et al., 2012)
		Direct Transfer	**Autoencoder SVM** (Zhou et al., 2014) **Adversarial** (Chen et al., 2018) **BLSE SVM** (Barnes et al., 2018)
Word Embeddings		Code-switched Monolingual	**Dict-CS** (Rasooli and Collins, 2017)
		Monolingual Mapping	**VecMap** (Artetxe et al., 2018) **MUSE** (Conneau et al., 2017)
		Bilingual-based	**BL** (Luong et al., 2015)
		Sentiment Embeddings	**BLSE** (Barnes et al., 2018)
Targeted	Monolingual	Aspect-based	**Feature Mining** (Hu and Liu, 2004) **Topic Modeling** (Brody and Elhadad, 2010) **Attention LSTM** (Wang et al., 2016)
		Entity-specific	**Target SVM** (Jiang et al., 2011) **Syntactic RNN** (Dong et al., 2014) **TC-LSTM** (Tang et al., 2015) **Attention LSTM** (Liu and Zhang, 2017)
		Open-domain	**Joint CRF** (Yang and Cardie, 2013) **PSL** (Deng and Wiebe, 2015a) **Integrated Neural** (Zhang et al., 2015)
		Other Languages	**Arabic Source** (Elarnaoty et al., 2012)
	Cross-lingual, Aspect-based	Annotation Projection	**Aspect Projection** (Klinger and Cimiano, 2015)
		Topic Models	**Aspect Model** (Zheng et al., 2014)
		Direct Transfer	**SMO** (Barnes et al., 2016) **biLSTM** (Akhtar et al., 2018)

Table 2.1: Summary of sentiment analysis work related to the thesis.

traditional techniques that rely on machine translation, models that try to incorporate unlabeled data from the target language, and cross-lingual models that transfer sentiment directly to the target language. As part of the discussion on cross-lingual sentiment analysis, we describe work on cross-lingual word and sentiment embedding representations in Section 2.2, and detail how they differ or relate to the word vector representations presented by this work.

We proceed to targeted sentiment analysis in Section 2.3, where we review the different formulations of the targeted sentiment task that have been presented in past studies, the techniques for annotating targets of sentiment as it relates to the dataset that we annotated, and the methods that have been studied for identifying targets and sentiment towards targets in open-domain, customer review analysis as well as shorter text formulations. We also survey methods where targeted sentiment analysis has been studied in languages other than English, focusing on language specific considerations that were learned in the process and motivating our work on Arabic targeted sentiment analysis.

Finally, in Section 2.4, we present related work on targeted cross-lingual sentiment analysis, which has been limited to a few studies. From this, we motivate our work on targeted cross-lingual sentiment analysis and describe how it differs from past work in the area. Table 2.1 shows a summary of related work as it compares to ours, with examples of notable work in each of the topics addressed.

2.1 Cross-lingual Sentiment Analysis

We describe past and contemporary work in cross-lingual sentiment analysis: methods that rely on machine translation, methods that rely on projection of annotations, and methods that transfer sentiment directly, or that use other means to transfer sentiment from a high-resource source language to a low-resource target language.

2.1.1 Machine Translation

The traditional approach to cross-lingual sentiment identification is to use a machine translation (MT) system. Machine translation methods either translate the target language text into a high-resource source language and apply a source-language model to predict the corresponding sentiment labels, or translate the source-language training data into the target language, from which a model can be trained in the target language.

The first approach was taken by Wan (2008), who use a publicly available machine translation systems to translate Chinese text to English and identify the sentiment using English sentiment lexicons. Additionally, Chinese sentiment lexicons are used to identify sentiment in the original Chinese reviews, and the two systems are combined in an ensemble to predict sentiment in Chinese. The Chinese-to-English approach outpeformed the in-language Chinese approach. Wan (2009) use a co-training approach: Chinese text is translated to English, and English training data is translated to Chinese, and a co-training algorithm is used to iteratively select the resulting data for classification using SVM models.

The work of Salameh et al. (2015) and Mohammad et al. (2016b) also explored both approaches: translating English training data to Arabic, and Arabic test data to English where a supervised English model predicts sentiment labels. Both manual and machine translation were employed, and better results were achieved on identifying sentiment by translating from Arabic to English rather than translating from English to Arabic, a conclusion that is supported by our cross-lingual Arabic-to-English and English-to-Arabic transfer experiments. Other studies, like that of Balahur and Turchi (2014), translate English training data into a number of European target languages - French, German, and Spanish - and build target-language sentiment models on the machine-translated text.

Machine translation approaches have also been developed using deep learning models. These include the work of Zhou et al. (2016), who translated the source training data into the target language and modeled both the source and target using a bidirectional LSTM, and Zhou et al. (2015), who translated the source training data in to the target language and used autoencoder models to create bilingual embeddings incorporating sentiment information from labeled data and their translations.

Machine translation-based solutions fall short, however, when the target language considered is a low-resource language that does not share a substantially large parallel corpus with the source language. The publicly available Google Translate[1], for example, is only available for about a hundred languages, but there are thousands of low-resource languages. Moreover, there are other problems with machine translation: it does not always preserve sentiment (Salameh et al., 2015), and it produces domain-mismatch in the vocabulary distributions in the original and translated data, with limited sentiment vocabulary in the translated target language data (Duh et al., 2011).

In this work, we take the position that a full-blown machine translation system is not necessary for the transfer of sentiment. Instead, we rely on bilingual representations of words, and bilingual representations of sentiment, that capture the context of sentiment expressed in both the source and target languages rather than only in the source language. Moreover, we provide the option of lexicalization, or partial, surface translation of the training data into the target language, for words that have translations in a bilingual dictionary. Because lexicalization does not apply any reordering or structural changes to the source text, this makes the translation less likely to alter the sentiment of the sentence.

[1] https://translate.google.com/

2.1.2 Annotation Projection

Annotation projection is similar to machine translation, except that it relies on making use of an existing parallel corpus rather than a machine translation system. Mihalcea et al. (2007) explored this approach by building a Romanian subjectivity classifier using annotation projections from a parallel English-Romanian corpus. A source-language model is first applied to the English side of the corpus, and the predicted sentiment labels are projected to the target language side. The projected labels can then be used to develop a sentiment classification model in the target language.

The advantage of the annotation projection method is that it does not rely on machine translation, and that it can make use of an out-of-domain corpus, if a high-resource, in-domain parallel corpus is not available. However, if such a translation corpus is available, it is faster and more advantageous to use the translation corpus to create bilingual representations that allow us to transfer sentiment directly using a single bilingual model, rather than separately training and running source and target language models.

2.1.3 Direct Sentiment Transfer

This work takes the approach of *direct* transfer of sentiment: a single cross-lingual model is trained on the source language, and subsequently applied to the target language. This mode of transfer is relatively new, but has been made possible by advances in machine learning and deep neural learning architectures (LeCun et al., 2015) and the development vector-based word representation, or word embedding, models (Turian et al., 2010), which are now possible to be trained in bilingual spaces (Hermann and Blunsom, 2013).

Direct transfer models have been applied to document classification (Upadhyay et al., 2016), named entity recognition (Täckström et al., 2012), parsing (Rasooli and Collins, 2015), as well as sentiment analysis. One such work is that of Zhou et al.

(2014), who used autoencoders to create shared sentence representations of English and Chinese from parallel data. Once shared bilingual sentence representations are learned, sentiment is classified using a simple SVM model that uses the sentence representations as input. This model uses only labeled data from the source language, and is thus a direct transfer model, but it allows the option of using labeled data from the target language. Additionally, experiments are performed on English-to-Chinese and Chinese-to-English transfer, making it one of the few studies that has used English as a target language. Our work is different because we rely on word-level bilingual representations and their associated sentiment features, and we extensively explore the role of different resources and feature representations amongst several language pairs.

The adversarial transfer model (Adv) of Chen et al. (2018) incorporates an adversarial training objective using a language predictor and a feature predictor. The language predictor tries to identify the language while the feature predictor tries to learn shared bilingual representations that are indistinguishable to the language predictor. They have several classifiers for making predictions: Deep Averaging Network (DAN), Convolutional Neural Network (CNN), Long Short-Term-Memory (LSTM), and LSTM with an attention mechanism; but they report CNN and LSTM with attention as their best performing models. Experiments are run using English as a source language and Chinese and Arabic as target languages. In theory, the approach does not require pre-trained word representations from translation corpora; however, the results presented by their work reveal that sentiment classification performance is much higher when pre-trained embeddings are used. Chapter 4 presents comparison results on our datasets using the adversarial model.

We also mention Barnes et al. (2018), work that is complementary to ours. This approach uses a single cross-lingual SVM model that relies on bilingual sentiment embeddings (BLSE). Unlike our work, it relies on projection matrices rather than translation corpora to create sentiment embeddings; and the sentiment embeddings and classifier are trained jointly, by using a bilingual sentiment lexicon to minimize the distance between source and target projection matrices. The size of the lexicon is assumed to be 10K-20K words. In contrast, the approach in this work relies on *pre-training* sentiment embeddings using an appropriately chosen translation corpus - parallel or comparable. The embeddings may optionally be updated during training, if target language lexicalization is applied.

2.1.4 Other Models

We mention other cross-lingual models that have avoided the use of MT, such as that of Meng et al. (2012), who used a cross-lingual mixture model. The motivation behind the work is similar to ours: to make use of a parallel corpus that directly identifies sentiment-carrying words in the target language, rather than identifying them by translating words from English. Their approach uses a cross-lingual mixture model (CLMM) to maximize the likelihood of generating a bilingual Chinese-English parallel corpus and determine word generation probabilities for each of the sentiment labels. Labeled data in the target language can but does not need to be available.

Unlike most previous work in cross-lingual sentiment analysis, the experiments in this thesis study direct transfer of sentiment using non-traditional resources, including comparable corpora and out-of-domain parallel corpora. To the best of our knowledge, this is the first work that uses comparable corpora as a translation resource for transferring sentiment; applications of comparable embeddings have been typically restricted to cross-lingual document classification or lexicon creation (Vulić and Moens, 2016). Moreover, unlike most of the work covered in this survey, the experiments in this work use moderately-resourced languages, including Arabic, Chinese and several moderately-resourced European languages, as not only target languages, but themselves also as source languages for transferring sentiment.

2.2 Word Embeddings

The success of direct sentiment transfer models depends to a large extent on the word vector representations that allow the model to be applied cross-lingually. Word embeddings allow lexical features to be represented in a continuous vector space, which captures not just the word but also its context. It is also possible to represent word vectors *cross-lingually* in a shared vector space occupied by different languages. Different approaches and resources have been proposed for training cross-lingual word vectors, but it is not clear which of these methods are best suited to the cross-lingual sentiment task.

In this section, we describe the main types of methods that have been used to train cross-lingual word embedding vectors, including methods for training word embedding vectors that incorporate sentiment (or sentiment embeddings). We refer to a number of these techniques throughout this work and particularly in Chapter 4, where we present detailed experimental analyses of the performance of monolingual-based (ML), bilingual-based (BL), and sentiment embeddings on the performance of our direct cross-lingual model.

2.2.1 Code-switched Monolingual Corpus (Dict-CS)

An efficient way to build word representations in multilingual spaces is to use a bilingual dictionary to code-switch, or partially translate words in monolingual data retrieved from different languages. Gouws and Søgaard (2015) and Rasooli and Collins (2017) used similar approaches to create code-switched, or mixed-language documents, on which monolingual word embedding models can be applied directly, such as that of Mikolov et al. (2013). This results in a *bilingual* word embedding space that consists of features from both source and target languages. We refer to this approach is dictionary code-switch, or DICT-CS. Implementing DICT-CS requires a large amount of monolingual data and a bilingual dictionary. If a manual bilingual dictionary is not available, it may be created automatically using word alignments from a parallel corpus, as in Rasooli and Collins (2017). We follow this approach in one of our cross-lingual models, which is presented in Chapter 4.

2.2.2 Mapping Monolingual Spaces (VecMap and MUSE)

In this approach, monolingual word vectors are induced separately in each language and a linear or non-linear projection is learned to map the vectors into the same space (Faruqui and Dyer, 2014; Ammar et al., 2016). The mapping relies on manual bilingual dictionary entries or on word alignments generated from parallel corpora.

More recently, Conneau et al. (2017) used a domain-adverserial setting and a refinement procedure that creates a synthetic dictionary that helps to further align the two language spaces (MUSE embeddings). Artetxe et al. (2018) proposed learning bilingual embeddings from only monolingual corpora by aligning monolingual embedding spaces by computing similarity matrices for each language and mapping the similarity matrices (VECMAP embeddings). Both approaches are advantageous and competitive if no translation corpora are available at all. However, it is unclear whether these approaches would outperform methods that use a small amount of translation data and directly utilize bilingual context. Furthermore, the monolingual embedding approaches above were developed and evaluated using corpora available for mostly European languages and it is less clear how they would perform with more low-resource languages.

Together, DICT-CS, VECMAP, and MUSE are considered to be monolingual-based (ML) word embedding models. Chapter 4 presents extensive experiments evaluating the performance of our direct cross-lingual models using each of these methods under varied conditions of resource availability.

2.2.3 Bilingual Embeddings from a Translation Corpus (BL)

Instead of using a monolingual corpus, this approach learns bilingual embeddings directly on a parallel corpus. We refer to this approach throughout the work as bilingual-based embeddings, or BL.

Luong et al. (2015) showed that learning bilingual embeddings directly on a parallel corpus produces embeddings that are high in both monolingual and bilingual quality. They propose a method to learn bilingual embeddings from a parallel corpus by extending the continous-bag-of-words and skip-gram (Mikolov et al., 2013).

In the monolingual Continuous-Bag-of-Words (CBOW) approach, the goal is to create word embedding representations by learning a language model that predicts a word w using its context c. Thus, given a "gold" corpus D containing pairs of words and contexts $(w, context)$, it models the probability that the observations $(w, context)$ occur in the data:

$$p(D = 1|w, context; \theta) = \frac{1}{1 + e^{-v_{context} \cdot v_w}} \tag{2.1}$$

where $v_w \in R^d$ is the word vector representation of w, $v_{context}$ is an average of context word vectors $v_c \in R^d$ in a window $\{-b, b\}$ around the center word w, and the softmax objective is maximized to learn parameters $\theta = v_w, v_c$ for all words and contexts.

The method of (Luong et al., 2015) extends the continous-bag-of-words (CBOW) and skip-gram (SG) models by learning bilingual models directly on the parallel corpora themselves. Word alignments are generated from the parallel corpora - although monotonic alignments can be assumed, which does not require learning a word alignment model. For each source or target word, both the monolingual context and bilingual contexts are used to predict it, essentially learning four joint models $s \rightarrow s$, $t \rightarrow t$, $s \rightarrow t$, $t \rightarrow s$ for source and target languages s and t. We build on this work in our approach and describe it in depth in Chapter 4.

2.2.4 Sentiment Embeddings

Here we describe work that incorporates sentiment into the representation of word vectors and point out how it differs from the sentiment embeddings proposed in this work. Similar to our work, Maas et al. (2011) predicted the sentiment of contexts in which a word occurs and used it in a word vector training objective. However, they annotated document-leve sentiment labels from online reviews and in a monolingual word vector setting, while we use only a source-language sentiment lexicon and learn our embeddings in a bilingual space. Similarly, Tang et al. (2016) combined word context with sentence-level sentiment evidence from a labeled sentiment dataset in their context-to-word prediction model, and Tang et al. (2014) used neural networks to learn sentiment embeddings from a distantly supervised Twitter dataset. However, all the above approaches focus on the monolingual space and use sentiment datasets or forms of distant learning to yield sentiment labels, while we use only a source-language sentiment lexicon.

Other work in the same vein includes that of Yu et al. (2017), who post-processed word vectors for sentiment by ranking nearest neighbors using a sentiment lexicon, and Faruqui et al. (2014), who refined word vectors in a post-processing step by using information from semantic lexicons.

On the other hand, there is less work that has explored sentiment embeddings bilingually. The approach of Zhou et al. (2015), referred to in Section 2.1.3, requires having translated sentiment labeled datasets available. The work of Barnes et al. (2018), referred to in Section 2.1.3, jointly learned bilingual sentiment embeddings with a sentiment classifier by using a bilingual sentiment lexicon to minimize the distance between source and target matrices. While their work assumes the availability of a large bilingual lexicon and uses projection matrices to learn embeddings, ours requires an existing parallel or comparable corpus, on which the embeddings are pre-trained, and optionally updated during training.

2.3 Targeted Sentiment Analysis

The targeted sentiment analysis task involves identifying sentiment towards a target, which could be an entity, a more generic topic, an aspect of customer service such as 'food' or 'ambiance', a product feature such as 'camera', or a situation such as the need for medical supplies during an earthquake. There are thus several formulations of the targeted sentiment task and several perspectives on approaches for annotating datasets with sentiment expressed towards targets. We survey these formulations along with the corresponding methods that have been studied, as well as approaches that have been taken for annotating datasets for targeted sentiment and how they differ from our work on annotating Arabic targeted sentiment in open-domain text.

2.3.1 Aspect-based

The earliest work in targeted sentiment analysis looked at identifying aspects and sentiment towards aspected in a restricted domain: that of product reviews or customer reviews. Many of these systems used unsupervised and topical methods for determining aspects of products; for example, Hu and Liu (2004) used frequent feature mining to find noun phrase aspects in product features, Brody and Elhadad (2010) used topic modeling to find important keywords in restaurant reviews, and Somasundaran and Wiebe (2009) mined the web to find important aspects associated with debate topics and their corresponding polarities. SemEval 2014 Task 4 (Pontiki et al., 2014) ran several subtasks for identifying aspect terms and sentiment towards aspects and terms in restaurant and laptop reviews, and SemEval 2016 Task 5 (Pontiki et al., 2016) involved identifying aspects in customer reviews in multiple languages, including Arabic, Chinese, Dutch, French, Russian, Spanish, and Turkish.

The work of Wang et al. (2016) integrated an attention mechanism into a long short-term-memory (LSTM) model in order to identify parts of the text that express sentiment towards aspects. Aspect embedding vectors are learned and concatenated with the word representation input at both the input and hidden state levels of the network. Our work in Chapter 7 uses a similar targeted attention mechanism, but towards targets instead of aspects, and in a cross-lingual model.

Aspect-based targeted sentiment differs from situation-based targeted sentiment, a task we introduce in this work, in several ways which we detail in Chapter 6.

2.3.2 Entity-specific

Entity-specific sentiment analysis involves identifying sentiment towards a target entity, e.g companies, politicians, or celebrities, and has typically been studied in genres of text that include social media and online posts. Jiang et al. (2011) proposed identifying sentiment of a tweet towards a specific named entity, taking into account multiple mentions of the given entity. SVM models as well as a graph-based sentiment optimization were used to take into account the different kinds of contextual tweets involving the target, such as retweets or tweets containing the target and published by the same person. Biyani et al. (2015) studied sentiment towards entities in longer online posts, similar to ours. In their study, the local part of the post that contained the entity or mentions of it was identified and the sentiment was classified using a number of linguistic features. The entities were selected beforehand and consisted of known, named entities. Our targeted sentiment work in Arabic on the other hand can freely identify any noun phrase as a target of sentiment.

Other work uses LSTM and RNN networks to determine sentiment towards entities in Twitter (Dong et al. (2014); Tang et al. (2015)). SemEval 2016 ran two tasks on sentiment analysis (Nakov et al., 2016) and stance (Mohammad et al., 2016a) towards topics in Twitter. Our SemEval Sentiment in Twitter Task 4 (Rosenthal et al., 2017) involved identifying both untargeted sentiment and sentiment towards topical entities in tweets, in both English and Arabic.

In earlier models that were developed, creating features for this task usually involved the heavy use of syntactic resources, e.g dependency parses, to determine the relationship between the target and nearby sentiment words (Jiang et al., 2011; Biyani et al., 2015; Somasundaran and Wiebe, 2009; Dong et al., 2014), which can be combined with deep learning methods (Dong et al., 2014); but with neural networks, such relationships may still be implicitly captured in the absence of syntactic resources. In the work of (Tang et al., 2015), for example, it was found that an LSTM-based targeted sentiment model outperformed other models that relied more heavily on syntax. In this approach, target embedding vectors are concatenated with input word embedding vectors at the input layer of the LSTM, and the sequence is split into left and right contexts with respect to the target, the output of which is concatenated before passing to a 'softmax' layer (TC-LSTM and TD-LSTM). On the other hand, the work of Liu and Zhang (2017) uses a target attention layer rather than a target embedding layer. Our work in Chapter 7 follows in this direction, except that we incorporate the targeted model cross-lingually.

2.3.3 Open-domain

Open-domain targeted sentiment analysis is similar to entity-based targeted analysis, but usually occurs in longer text that involves multiple targets of sentiment per post, and is generally not restricted to named entity targets. In early work, Kim and Hovy (2006) proposed finding opinion target and sources in news text by automatic labeling of semantic roles. Here, opinion-target relationships were restricted to relations that can be captured using semantic roles. Ruppenhofer et al. (2008) discussed the challenges of identifying targets in open-domain text which cannot be addressed by semantic role labeling, such as implicitly conveyed sentiment, global and local targets related to the same entity, and the need for distinguishing between entity and proposition targets.

Sequence labeling models became more popular for this problem: Mitchell et al. (2013) used CRF model combinations to identify named entity targets in English and Spanish, and Yang and Cardie (2013) used joint modeling to predict opinion expressions and their source and target spans in news articles, improving over several single CRF models. Their focus was on identifying directly subjective opinion expressions (e.g "I *hate* [this dictator]" vs. "[This dictator] is *destroying* his country.") The work of Deng and Wiebe (2015a), which is based on probabilitstic soft-logic models (PSL), identifies entity sources and targets, as well as the sentiment expressed by and towards these entities. This work was based on probablistic soft logic models, also with a focus on direct subjective expressions. In the same spirit, our work in open-domain Arabic targeted sentiment uses CRF models, but we do not identify sources of sentiment or the sentiment expressions themselves.

There has also been work on using neural networks for tagging open-domain targets (Zhang et al., 2015; Liu et al., 2015) in shorter posts. In contrast to our work, previous work listed did not consider word morphology, or explicitly model distributional entity semantics as indicative of the presence of sentiment targets. Our work on open-domain targeted sentiment models morphological representation features as well as discrete cluster embedding features. It has been shown consistently that semantic word clusters improve the performance of named entity recognition (Täckström et al., 2012; Zirikly and Hagiwara, 2015; Turian et al., 2010) and semantic parsing (Saleh et al., 2014); motivated by this success, we use clusters in addition to morphological representation features in our models for identifying entity targets of sentiment in Arabic.

2.3.4 Other Languages

Studies on targeted sentiment analysis are less prevalent in other languages compared to English. Examples of targeted sentiment study in other languages include that of Elarnaoty et al. (2012), who proposed identifying sources of opinions in Arabic using a conditional random field (CRF) with a number of patterns, lexical and subjectivity clues; in contrast to our work, they did not discuss morphology or syntactic relations. Al-Smadi et al. (2015) developed a dataset and built a majority baseline for finding targets in Arabic book reviews of known aspects; Obaidat et al. (2015) also developed a lexicon-based approach to improve on this baseline. Abu-Jbara et al. (2013) created a simple opinion-target system for Arabic by identifying noun phrases in polarized text; this was done intrinsically as part of an effort to identify opinion subgroups in online discussions. Ours is the earliest work (Farra and McKeown, 2017) to use sequence labeling models to identify target entities and sentiment in Arabic or to use open-domain text, but other work that uses sequence labeling, deep learning, or some morphological features has followed since then for Twitter (El-Kilany et al., 2018) and book review (Al-Smadi et al., 2018) genres.

Past work in Arabic machine translation (Habash and Sadat, 2006) and named entity recognition (Benajiba et al., 2008) considered the tokenization of complex Arabic words, but analysis of such segmentation schemes has not been previously reported for Arabic sentiment tasks, which have typically covered mostly untargeted sentiment analysis with lemma or surface bag-of-word representations. In our work, however, we consider the impact of morphological-based segmentation on both Arabic targeted sentiment analysis as well as on our cross-lingual models that use Arabic as a source language.

There has also been previous work on identifying targeted sentiment in Chinese, which includes Lipenkova (2015) who used unit identification and relation extraction models for aspect-based sentiment analysis in Chinese, and Peng et al. (2018), who proposed a formulation of the aspect-based targeted sentiment task that is more suitable for the characteristics of Chinese linguistics, such as sub-element characters. Syed et al. (2014) identified sentiment towards targets in Urdu, a morphologically complex language, using shallow parse chunking, dependency relations, and Urdu sentiment lexicons. Other notable targeted sentiment datasets include the multilingual aspect-based SemEval 2016 Task 5 dataset (Pontiki et al., 2016).

2.3.5 Annotating Targets

Here we review previous work in target annotation in English and Arabic, describing how it differs from our work on creating an open-domain targeted dataset for Arabic.

2.3.5.1 Annotating Targets in English

One of the early datasets collected for identifying sentiment targets is that of Hu and Liu (2004), where product features (e.g price, quality) were annotated in customer reviews of consumer electronics. These consisted of mostly explicit product features annotated by one person. Also in the product review domain, the SemEval 2014 task of Pontiki et al. (2014) was concerned with finding aspect categories of products along with the sentiment expressed towards them. The products (e.g 'restaurant') and coarse-grained features (e.g 'service') were provided to annotators, who identified the aspect terms (e.g 'waiter') and the corresponding sentiment expressed towards them.

The MPQA corpus is an in-depth and general-purpose resource for fine-grained subjectivity and sentiment annotations (Wiebe et al., 2005; Wilson, 2008), containing annotations of sentiment expressions at the phrase level while specifying polarities, sources, and target spans. The annotation scheme links each subjective expression to one or more attitudes, which in turn can have one or more or no targets. The target annotations include the full target spans, but do not necessarily identify target entities within the span. Stoyanov and Cardie (2008) extended part of the MPQA corpus by annotating it for 'topics', arguing that 'targets' refer to the syntactic span of text that identifies the content of a sentiment expression, while 'topic' is the real-world object or entity corresponding to the primary subject of the sentiment expression. Using trained annotators, they identify 'topic clusters', which group together all opinions referring to the same topic. In parallel with this work, part of the MPQA corpus was annotated for entity-level targets (Deng and Wiebe, 2015b) by specifying target entities within the MPQA span, leading to the annotation of 292 targets by two annotators. The entities were anchored to the head word of the noun phrase or verb phrase that refers to the entity or event. In our work, which includes sentiment annotations of 4345 target entities, we only consider noun phrase entities, and we consider the noun phrase itself as an entity.

Other target annotation studies include that of Toprak et al. (2010) who enrich target and source annotations in consumer reviews with measures such as relevancy and intensity, and Somasundaran et al. (2008) who perform discourse-level annotation of opinion frames, which consist of sentiment expressions whose targets are described by similar or contrasting relations. In most of these studies, the annotation was usually done by trained individuals or someone who has knowledge and experience in the task. Our work on annotating Arabic is different in that it utilizes crowdsourcing for the annotation process, and it focuses on the marking of important entities and concepts as targets of sentiment expressions in the more genre of online comments to newspaper articles. We view targets as 'real-world entities', similar to the topics discussed by Stoyanov and Cardie (2008), and the targets in Deng and Wiebe (2015b), and we annotate multiple targets in the text.

Carvalho et al. (2011) also annotated targets in online comments; here targets were considered to be human entities, namely political and media personalities. This annotation was done by one trained annotator where agreement was computed for a portion of the data. Another related task was that of Lawson et al. (2010) who describe a Mechanical Turk annotation study for annotating named entities in emails, with favorable agreement results. The tasks for identifying the spans of and labeling the named entities were grouped in a single Human Intelligence Task (HIT).

2.3.5.2 Annotating Sentiment in Arabic

Besides the aspect-based book review dataset of Al-Smadi et al. (2018), the targeted work mentioned in Section 2.3.4, and our own work on collecting a targeted sentiment dataset for SemEval 2017 Task 4 (Rosenthal et al., 2017), we mention work on annotation of untargeted sentiment in Arabic, which includes the sentence-level annotation study of Abdul-Mageed and Diab (2011) for Modern Standard Arabic (MSA) newswire data, and which covers multiple domains including politics, sports, economy, culture, and others; both the domains and the sentence-level sentiment were annotated by two trained annotators. Our Arabic annotation data also comes from different domains, but it is from the genre of online comments to newspaper articles, which have greater prevalence of dialect, imperfect grammar, and spelling errors.

There have been other crowdsourcing annotation studies in Arabic; among them Zaidan and Callison-Burch (2011) who annotated dialectness, Denkowski et al. (2010) who annotated machine translation pairs, and Higgins et al. (2010) who annotated Arabic nicknames.

2.4 Targeted Cross-lingual Sentiment Analysis

In our study of the field, we have encountered very little previous work that studied cross-lingual targeted sentiment analysis. We mention here the few published systems we have found, all of which involve aspect-based sentiment analysis and most of which do not integrate a target-specific modeling mechanism in the model.

The contemporary study that is most related to our work is that of Akhtar et al. (2018), who developed a direct cross-lingual sentiment model for aspect-based sentiment analysis in English-Hindi and English-French, using a bidirectional Long Short-Term-Memory (biLSTM) model. They employ an approach similar to target language lexicalization, which addresses out-of-vocabulary (OOV) words by translating into English and mapping back to an in-vocabulary target language word. Their English-Hindi in-domain parallel corpus used 7.2 million sentences generated using an MT system to learn bilingual embeddings, which is not a truly low-resource setting; on the other hand, our largest in-domain parallel corpus (for all experiments, untargeted or targeted) is less than 400K sentences and our smallest is 11K sentences, while our largest out-of-domain parallel corpus is 860K sentences. They did not employ any target-specific modeling - i.e, an untargeted model is used for targeted sentiment identification - and while we create bilingual sentiment features whose weights may be updated in a bilingual space, they rely on projecting sentiment scores from an English lexicon.

Another approach (Barnes et al., 2016) used direct cross-lingual models using Sequential Minimal Optimization (SMO) classifiers to identify aspect-based sentiment analysis in English and Spanish. The input to their model was not a sentence, but opinion units (source, target, and sentiment expression) that are already known, for which the aspect-based sentiment is to be determined. They do not integrate any target-specific mechanism in their model. They compare bilingual-based embedding models, projection-based monolingual embedding models, stacked auto-encoders models, and high-resource machine translation models that translate the opinion units in context. Their bilingual-based embedding models outperform all but the high-resource MT models trained on in-domain data, and they furthermore found that the performance of bilingual-based embeddings were stable with parallel data size. Their results are consistent with our own; our experiments our more expansive, evaluating a number of monolingual-based and bilingual-based methods, bilingual sentiment embeddings, as well as smaller parallel corpora sizes and out-of-domain corpora.

A different kind of approach was taken by Klinger and Cimiano (2015), which aims to detect aspect (i.e, target) phrases themselves along with the sentiment expressions in the target language. To do this, they use machine translation to translate the source (English) training data into the target (German) language, and then project the annotation of target words using word alignments, where a model can be trained in the target language. They improve the performance of their model by filtering the sentences selected for projection based on machine translation quality, assessed using language models.

Finally, we mention Zheng et al. (2014), who used cross-lingual topic modeling in order to jointly identify aspects along with their sentiment in hotel reviews taken from a number of languages including English, Chinese, French, German, Spanish, Dutch, and Italian.

Of the approaches mentioned, all have focused on the problem of aspect-based sentiment analysis, and none has used a cross-lingual model with a mechanism that incorporates sentiment towards targets. Our targeted cross-lingual model is assessed on identifying sentiment towards entities in Twitter, incorporates an attention mechanism towards targets while incorporating bilingual sentiment embeddings, and is evaluated on English-to-Arabic *as well as* Arabic-to-English cross-lingual targeted sentiment transfer.

Chapter 3

Resources for Cross-lingual Sentiment Analysis

> **"There is no deficit in human resources; the deficit is in human will."**
>
> — Martin Luther King Jr., *Nobel Lecture, Dec. 1964*

To identify sentiment in a low-resource language, we must begin with identifying the resources available for our language. Depending on what kind of cross-lingual resources are available for training system features - size, quality, and suitability for the study of sentiment - our system can yield very different results.

With low-resource languages, cross-lingual resources are often scarce. While parallel translation corpora for some of the more high-resource European languages is available in the order of millions of sentence pairs, resources like large parallel corpora, bilingual dictionaries, and sentiment lexicons are often not available for languages like Uyghur, a low-resource Turkic language spoken by the Muslim minority in the Xinjiang region of China. Instead, to transfer sentiment effectively to low-resource languages, we must find alternative resources, rely more heavily on monolingual data, or find ways to effectively utilize smaller parallel corpora. For this reason, our work focuses on collecting resources that have not been traditionally used in cross-lingual sentiment analysis tasks; these include religious corpora, comparable corpora, and evaluation datasets for newly selected low-resource languages whose cross-lingual sentiment performance has not been previously studied. In contrast, most previous work in the field has relied on using large in-genre parallel corpora (such as those used for building machine translation systems) and resources that are typically only available for European and Indo-European languages.

Moreover, the performance of the cross-lingual system is likely to be impacted by the choice of the source and target languages themselves: whether they are syntactically and semantically similar, or in the same language family. Thus, the choice of source and target languages is another important consideration that will be addressed in this chapter as well as later ones.

Training data is yet another highly important resource: while our training and evaluation data will inevitably come from two completely different languages, we can at least select the genres and domains of our training data to be similar to that of our evaluation data. The collection of cross-lingual resources in our work includes training data for targeted and untargeted sentiment analysis, bilingual corpora used for transfer, and metrics for studying the comparability of these corpora.

This chapter describes efforts to build cross-lingual resources for sentiment analysis, starting with the identification of appropriate source and target languages for analysis, the collection of training and evaluation data used for untargeted and targeted sentiment analysis, followed by the collection of untraditional resources for cross-lingual sentiment analysis, which include both parallel and non-parallel corpora, and the study of the comparability of these bilingual corpora.

In Section 3.1, we describe the source and target languages studied throughout the work.

In Section 3.2, we discuss native informants and their role in the annotation of evaluation datasets.

In Section 3.3, we introduce the untargeted and targeted sentiment datasets that will be used throughout the work. We have produced three sentiment analysis datasets for Arabic as part of the thesis: two Twitter datasets consisting of targeted and non-targeted annotations created in conjunction with our work on organizing SemEval-2017 (Rosenthal et al., 2017), and a targeted sentiment dataset of news article comments (Farra et al., 2015a). We introduce the last dataset in this chapter and detail its collection in Chapter 5, where we discuss open-domain targeted sentiment analysis.

In Section 3.4, we describe the parallel data resources used in the work, which include data from the Linguistic Data Consortium, the European Parliament corpus, and the Bible and Quran.

In Section 3.5, we introduce the comparable corpora we created, which include topic-aligned and document-aligned Wikipedia corpora collected for 18 languages.

In Section 3.6, we describe our monolingual corpora, which include Wikipedia dumps as well as monolingual data provided by the Linguistic Data Consortium.

Finally, in Section 3.7, we introduce and compute two metrics for measuring the comparability of the bilingual parallel and comparable resources, translation comparability (Li and Gaussier, 2010) and our extension for computing sentiment comparability.

3.1 Languages

Table 3.1 shows the languages considered in the thesis. They are divided among six language families: Afro-Asiatic, Sino-Tibetan, Uralic, Turkic, and Indo-European. The Indo-European languages are divided among five sub-families: Indo-Iranian, Indo-Aryan, Slavic, Romance, and Germanic.

In selecting our languages, we included both target languages that are truly *low-resource* - i.e, where virtually no training data, NLP systems, or tools exist for the purpose of sentiment analysis, e.g Tigrinya and Uyghur, as well as more highly resourced languages such as Spanish or Arabic, where it is easier to acquire larger evaluation datasets or online machine translation systems that facilitate the error analysis for our cross-lingual models.

Language	Code	Family	Sub-Family
Arabic	ar	Afro-Asiatic	Semitic
Tigrinya	ti		
Sinhalese	si	Indo-European	Indo-Aryan
Persian	fa	Indo-European	Indo-Iranian
English	en	Indo-European	Germanic
German	de		
Swedish	sv		
Spanish	es	Indo-European	Romance
Portuguese	pt		
Bulgarian	bg	Indo-European	Slavic
Croatian	hr		
Polish	pl		
Russian	ru		
Slovak	sk		
Slovene	sl		
Mandarin Chinese	zh	Sino-Tibetan	Sinitic
Uyghur	ug	Turkic	Karluk
Hungarian	hu	Uralic	Finno-Ugric

Table 3.1: Languages, families and sub-families considered in the thesis. The second column represents the ISO 639-1 language code Byrum (1999).

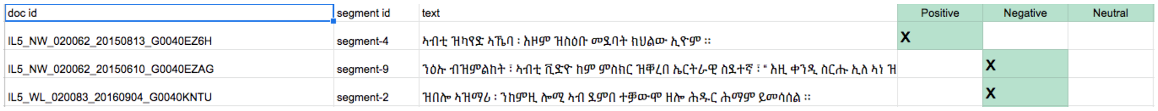
In addition, nine of our target languages have been included as either representative languages (Arabic, Hungarian, Russian, Persian, and Spanish) or incident languages (Chinese, Uyghur, Tigrinya, and Sinhalese) for the DARPA Low Resource Languages for Emergent Incidents (LORELEI) low resource language program (Christianson et al., 2018). The program generally selects languages which have significant numbers of native speakers but are less represented in terms of language resources (Cieri et al., 2016). Our remaining languages, with the exception of Croatian, are European Parliament (EU) languages, some of which may themselves (e.g Slovak) be considered as low resource or moderately resourced languages (Maxwell and Hughes, 2006).

We treat languages alternatively as source or target languages during cross-lingual transfer of sentiment. The Indo-European languages (with the exception of Sinhalese) along with Arabic and Chinese[1], will alternatively play the role of both source and target languages. Because of the lack of training data, Sinhalese, Tigrinya, Uyghur will be treated only as target languages.

3.2 Native Informants

When resources for a language are rare, having access to even an hour of a native speaker's time can be a valuable source of information. While developing our language resources, we asked for assistance from native language informants. The informants were asked to complete the annotation tasks remotely on a web interface. The amount of time spent by any informant on providing us with information on a given language was limited to 60 minutes. We made use of knowledge from native informants in the following ways:

1. Annotation of sentiment evaluation datasets on a three-point scale for Uyghur, Tigrinya, Sinhalese, and Chinese[2]. Figure 3.1 shows an example of the sentiment annotation setup.



doc id	segment id	text	Positive	Negative	Neutral
IL5_NW_020062_20150813_G0040EZ6H	segment-4	ኣብቲ ዝካየድ ኣኼባ ፡ እዞም ዝስዕቡ መደባት ክህልው ኢዮም ።	X		
IL5_NW_020062_20150610_G0040EZAG	segment-9	ንዕኡ ብዝምልከት ፧ ኣብቲ ቪድዮ ከም ምስክር ዝቐረበ ኤርትራዊ ስደተኛ ፧ " እዚ ቀንዲ ስርሑ ኢሉ ኣነ ዝ		X	
IL5_WL_020083_20160904_G0040KNTU	segment-2	ዝበሎ ኣዝማሪ ፡ ንከምዚ ሎሚ ኣብ ደምበ ተቓውሞ ዘሎ ሕዱር ሕማም ይመሳሰል ።		X	

Figure 3.1: Example of native informant annotation interface.

2. For Tigrinya, the native informant was asked to manually translate keywords from English to the target language, and the keywords were subsequently used for collecting our comparable corpora as will be described in subsequent sections. To save time, the annotator was provided with keyword translations for words that we had found in online dictionaries, and was asked to verify the

[1] Modern Standard Mandarin Chinese will be referred to as Chinese throughout the work.

[2] The Chinese language informant was a graduate student who used a different setup and was only asked to perform this first task.

translations. Online dictionaries and Google Translate were used to translate these keywords for the other languages.

3.3 Sentiment Datasets

We describe here the training and evaluation data for our cross-lingual untargeted and targeted sentiment tasks. Table 3.2 shows a summary. The datasets for the Open Domain and Situation Frame targeted tasks, among them the Open Domain sentiment dataset we collected for Arabic, will be described in depth in Chapter 4 when we introduce the Open Domain and Situation Frame targeted tasks.

3.3.1 Untargeted Datasets

The untargeted datasets consist of text annotated for sentiment at the sentence level with one of three labels $l \in \{positive, negative, neutral\}$. Training data - albeit variant in size - is available for all languages except Uyghur, Tigrinya, and Sinhalese.

3.3.1.1 European Twitter Dataset

For the twelve European languages, we have used the tweets downloaded from the Twitter dataset of Mozetič et al. (2016). The datasets for each language are split into 80% train, and 10% test, leaving 10% aside for development data. The dataset sizes, along with their distribution amongst sentiment labels, are shown in Table 3.3. The train and test sets are similarly distributed for sentiment.

3.3.1.2 Persian Product Reviews

The dataset for Persian was obtained from the SentiPers data (Hosseini et al., 2015), a set of digital product reviews. The dataset was split into 80% train and 10%

Task	Dataset	Source	Training	Evaluation
Untargeted	European Twitter	Mozetič et al. (2016)	bg, de, es, en, hu, hr, pl, pt, ru, sk, sl, sv	bg, de, es en, hu, hr, pl, pt, ru, sk, sl, sv
	SemEval-2017 A	Rosenthal et al. (2017) (our collaborative work)	ar	ar
	Syria Dataset	Salameh et al. (2015)	ar	–
	BBN Dataset	Salameh et al. (2015)	ar	–
	SAMAR	Abdul-Mageed et al. (2014)	ar	–
	Product Reviews	Hosseini et al. (2015)	fa	fa
	Hotel Reviews	Lin et al. (2015)	zh	zh
	Mono-LDC IL2	Native Annotated	–	zh
	Mono-LDC IL3	Native Annotated	–	ug
	Mono-LDC IL5	Native Annotated	–	ti
	Mono-LDC IL10	Native Annotated	–	si
Targeted	Dong Twitter	Dong et al. (2014)	en	en
	SemEval-2017 B,C	Rosenthal et al. (2017) (our collaborative work)	en, ar	en, ar
	Open Domain	Farra et al. (2015a) (this work)	ar	ar
	Situation Frame	LDC	en, es	en, es

Table 3.2: Untargeted and targeted sentiment datasets with training and evaluation languages.

	bg	de	en	es	hu	hr	pl	pt	ru	sk	sl	sv
Train	23739	63669	46622	137106	36167	56212	116105	62989	44757	40470	74238	32600
% P	28.9	25.6	28.9	48.4	51.8	53.2	43.6	27.3	28.0	54.0	26.2	26.6
% N	20.3	18.2	24.9	11.0	14.9	23.8	30.3	38.1	30.5	24.7	28.6	42.2
% O	50.8	56.2	46.3	40.6	33.2	23.0	26.2	34.6	41.5	21.3	45.3	31.2
Test	2958	7961	5828	17133	4520	7025	14517	7872	5594	5058	9277	4074

Table 3.3: European Twitter dataset with train and test size in sentences, and distribution amongst sentiment labels (P:positive, N:negative,O:neutral).

test, leaving aside 10% for development data. The dataset sizes, along with their distribution amongst sentiment labels, are shown in Table 3.4.

Persian Product Reviews	fa
Train	15000
% P	50.3
% N	10.2
% O	39.6
Test	3027
% P	52.8
% N	10.3
% O	36.8

Table 3.4: Persian product reviews dataset with train and test size in sentences, and their distribution amongs sentiment labels (P:positive, N:negative, O:neutral).

3.3.1.3 SemEval-2017 Dataset

The SemEval task on Sentiment Analysis in Twitter task has been run multiple times since 2014 (Rosenthal et al., 2014, 2015b, 2017) and has included subtasks and benchmark datasets for untargeted and targeted sentiment prediction. SemEval-2017 Task 4 (Rosenthal et al., 2017), which we co-organized, included an Arabic benchmark dataset for the first time. This dataset was created through our contribution.

The untargeted Arabic sentiment dataset was collected for SemEval2017 Task 4 Subtask A, where the goal is to predict untargeted sentiment in three categories: positive, negative, and neutral.

3.3.1.4 Other Arabic Training Datasets

In addition to the Arabic SemEval dataset, and in order to acquire a sufficiently large amount of training data for Arabic, we collected additional sources of Arabic training data for untargeted sentiment analysis. They are as follows: the Syria dataset of Salameh et al. (2015) consisting of tweets originating from Syria (where the Levantine dialect of Arabic is commonly spoken) polled from Twitter in May 2014, the BBN dataset of Salameh et al. (2015), a subset of the BBN Arabic-Dialect/English Parallel text corpus (Zbib et al., 2012) randomly selected for sentiment annotation, and the SAMAR Twitter dataset of Abdul-Mageed et al. (2014). The texts in the

SAMAR dataset are annotated for labels ‘positive’, ‘negative’, ‘neutral’ ,‘objective, and ‘mixed’; we combined ‘objective’ and ‘neutral’ into a single category and omitted ‘mixed’ labels. After processing for errors, the entire Arabic training data for untargeted sentiment amounted to 8385 sentences.

We utilized these datasets fully for training purposes and kept the SemEval test set, which is relatively large in size, as our benchmark Arabic evaluation data. Table 3.5 shows the breakdown of the complete Arabic untargeted train and test sets along with the dataset sizes.

Arabic Training Data	SemEval-2017 A	Syria	SAMAR	Total
Train	2684	2000	2503	8385
% P	19.4	22.4	17.5	22.7
% N	37.8	67.5	28.4	43.5
% O	42.8	10.1	53.9	33.7
Test	6100	–	–	6100
% P	24.8	–	–	24.8
% N	36.4	–	–	36.4
% O	38.8	–	–	38.8

Table 3.5: Arabic untargeted sentiment datasets with train and test sizes, and distribution amongst sentiment labels (P:positive, N:negative, O:neutral).

3.3.1.5 Chinese Datasets

For Mandarin Chinese, we have used training data from the Hotel Reviews dataset of Lin et al. (2015) for running experiments with Chinese as a source language. The dataset consists of 170K hotel reviews annotated for sentiment on a 5-point scale and balanced amongst the 5 classes, whereby we have consolidated all positive or negative classes to create a 3-point dataset. For the Chinese target evaluation data, however, we used a subset of the monolingual language pack provided by LDC[3] for Chinese, an incident language in the LORELEI program. This evaluation dataset was annotated by a native speaker. Table 3.6 shows the sizes of the Chinese evaluation dataset.

[3] LDC2016E30_LORELEI_Mandarin_Incident_Language_Pack_V2.0

	zh (IL2)
Mono-LDC IL2 Test	487
% P	27.7
% N	30.2
% O	42.1

Table 3.6: Chinese evaluation datasets with test size in sentences, and distribution amongst sentiment labels (P:positive, N:negative, O:neutral).

3.3.1.6 Languages with No Training Data

We created evaluation datasets using subsets of the monolingual data supplied by LDC as part of its LORELEI incident language packs for low-resource languages. As mentioned in Section 3.2, we annotated these datasets by relying on the help of native speakers of these languages. The number of sentences annotated was thus limited by the availability of the native informant, which was time restricted. Table 3.7 shows the evaluation dataset sizes and sentiment distribution for Mono-LDC datasets for Uyghur, Tigrinya, and Sinhalese, where we do not have access to any sentiment training data.

	si (IL10)	ti (IL5)	ug (IL3)
Mono-LDC Test	295	239	346
% P	25.4	10.9	19.1
% N	42.7	36.0	26.0
% O	31.9	53.1	54.9

Table 3.7: Evaluation set sizes for Mono-LDC IL3, IL5, IL10 for languages with no training data (P:positive, N:negative, O:neutral).

We note that all the incident languages, as well as Arabic (Tables 3.5, 3.6, and 3.7) have relatively high occurence of negative sentiment labels in the evaluation data. The sentiment distribution of many of the training languages; however, has a higher occurrence of neutral and positive data (Table 3.3). This mismatch in distribution of sentiment labels in fact mirrors the real life situation of many cross-lingual sentiment applications, where training data has a natural distribution of high neutral and positive occurrence but evaluation data that occurs in the case of a disaster incident contains a much higher occurrence of negative sentiment. Our approach to cross-lingual sentiment transfer will address this disparity by pre-training our cross-lingual models with bilingual sentiment features learned on corpora that are high in sentiment content. These bilingual features help detect negative sentiment in the target language even when the source language training data is biased towards neutral or positive sentiment.

3.3.2 Targeted Datasets

Here we introduce the datasets for targeted sentiment analysis. Of these datasets, two result from our own work, the SemEval-2017 Arabic Targeted Dataset with the newly collected English Targeted Test Dataset, and the Arabic Open Domain Dataset.

3.3.2.1 SemEval-2017 Targeted Dataset

The SemEval-2017 targeted dataset was collected for Task 4 Subtasks B and C, where the goal is to predict targeted sentiment towards topics in two categories (positive and negative: Subtask B), and five categories (highly positive, weakly positive, neutral, weakly negative, and highly negative: Subtask C). We have participated in the collection of this dataset, particularly the Arabic train and test datasets, which were newly added to the task in 2017.

The dataset was collected by scraping tweets mentioning a number of different trending topics, internationally and in Arabic-speaking parts of the world, using local and global Twitter trends[4] and it was annotated for sentiment *towards* the topics using the annotation platform CrowdFlower, now known as Figure Eight[5]. In doing so we stressed that targeted positive or negative sentiment needed to express

[4] https://trends24.in

[5] https://www.figure-eight.com/

Tweet	Untargeted Sentiment	Targeted Sentiment
Who are you tomorrow? Will you make me smile or just bring me sorrow? #HottieOfTheWeek Demi Lovato	NEUTRAL	Demi Lovato: POSITIVE
Saturday without Leeds United is like Sunday dinner it doesn't feel normal at all (Ryan)	WEAKLYNEGATIVE	Leeds United: HIGHLYPOSITIVE
Apple releases a new update of its OS	NEUTRAL	Apple: NEUTRAL

Table 3.8: Some English example annotations that we provided to the annotators.

Tweet	Untargeted Sentiment	Targeted Sentiment
أبل تطلق النسخة التجريبية الرابعة لنظام التشغيل *Apple releases a fourth beta of its OS*	NEUTRAL	أبل *Apple*: NEUTRAL
المايسترو … الاسطورة روجر فدرر ملك اللّعب الخلفي من اجمل لقطاته *The maestro … the legend Roger Federer king of the backhand game one of his best shots*	HIGHLYPOSITIVE	فدرر *Federer*: HIGHLYPOSITIVE
اللّاجئون يواجهون الصعوبات *Refugees are facing difficulties*	WEAKLYNEGATIVE	اللّاجئون *Refugees* : NEUTRAL

Table 3.9: Some Arabic example annotations that we provided to the annotators.

an opinion about the topic itself rather than about a positive or a negative event occurring in the context of the topic (see for example the third row of Table 3.9).

The topics included a range of named entities (e.g., *Donald Trump*, *iPhone*), geopolitical entities (e.g., *Aleppo*, *Palestine*), and other entities (e.g., *Syrian refugees*, *Dakota Access Pipeline*, *Western media*, *gun control*, and *vegetarianism*). We then used the Twitter API to download tweets, along with corresponding user information, containing mentions of these topics in the specified language. We intentionally chose to use some overlapping topics between the two languages in order to encourage cross-lingual approaches.

Tables 3.8 and 3.9 show examples of the tweets along with their targeted and untargeted annotations. In our experiments, we consolidate weakly negative and weakly positive annotations with highly negative and highly positive ones.

Figure 3.10 shows the statistics of the Arabic and English targeted datasets. We observe that for Arabic there is a higher distribution of neutral labels and less negative labels with the same dataset compared to the untargeted annotations.

More information about the details of the data collection and annotation can be found in our SemEval task paper (Rosenthal et al., 2017).

SemEval Targeted Datasets	English	Arabic
Train	20,508	3355
# Annotated Topics	125	34
% P	72.9%	26.4%
% N	19.7%	23.0%
% O	7.5%	50.6%
Test	12,379	6100
# Annotated Topics	125	61
% P	19.9%	25.6%
% N	30.1%	19.6%
% O	50.0%	54.8%

Table 3.10: SemEval 2017 English and Arabic targeted sentiment datasets with train and test sizes, number of topics, and distribution amongst sentiment labels (P:positive, N:negative, O:neutral).

3.3.2.2 Dong Dataset

Because the English SemEval targeted training data is heavily biased towards positive sentiment, we also use an additional Twitter dataset, collected and manually annotated by Dong et al. (2014) for targeted sentiment analysis in English. This dataset consists of 6,248 train tweets and 692 test tweets, divided amongs neutral, positive, and negative samples in proportions 50%, 25%, and 25% respectively.

3.3.2.3 Arabic Open Domain Dataset

The Arabic Open Domain Dataset consists of online comments to *Aljazeera* news articles annotated for targeted sentiment towards multiple entities in multiple domains: politics, culture, and sports. It was collected in our work on annotation of targeted sentiment using crowdsourcing (Farra et al., 2015a). Because the description of the dataset collection is tied to the nature of the Open Domain targeted task rooted in longer document-style comments, we describe this dataset in detail in Chapter 6.

3.3.2.4 Situation Frame Dataset

The Situation Frame Dataset consists of news, Twitter, and discussion forum documents, annotated by LDC for targeted sentiment towards *situation frames* and entities. Similarly to the open domain dataset, because the situation frame task is a new problem that is different than the traditional targeted sentiment task, and is rooted in longer document-style text, we describe this dataset in detail in Chapter 6.

3.4 Parallel Corpora

The availability of parallel translation corpora, providing human annotated translations between languages at the sentence level, is a central component of many cross-lingual natural language processing systems, among them machine translation systems. However, machine translation systems require very large parallel corpora, sometimes on the order of millions of sentences, in order to effectively model phrasal alignments and generate syntactically correct translated text for the target language. Most low-resource languages lack this translation data. Moreover, for cross-lingual classification of sentiment labels, such large parallel corpora may not be necessary. A sufficient amount of parallel data is only needed to achieve the following:

- **Bilingual Feature Representations**: Generate bilingual feature representations, or word representations in a bilingual space, using the parallel corpora.
- **Bilingual Dictionaries**: Generate bilingual dictionaries, if needed by the model, by generating word alignments using the parallel corpora.

We collected parallel corpora from a number of different sources, including untraditional sources such as texts from the Bible and Quran, with the goal of studying the effect of the genre and size of the corpus on the performance of cross-lingual sentiment analysis in the target language. We describe here the parallel corpora we used: (1)

'High-quality' in-domain parallel corpora from the LDC, (2) Contemporary parallel corpora from the European Parliament containing political text, and (3) Religious parallel corpora combined using translations from the Bible and Quran.

3.4.1 Linguistic Data Consortium

The most ideal scenario for cross-lingual sentiment occurs when we have large amounts of parallel data that are *in-genre* and *in-domain*.

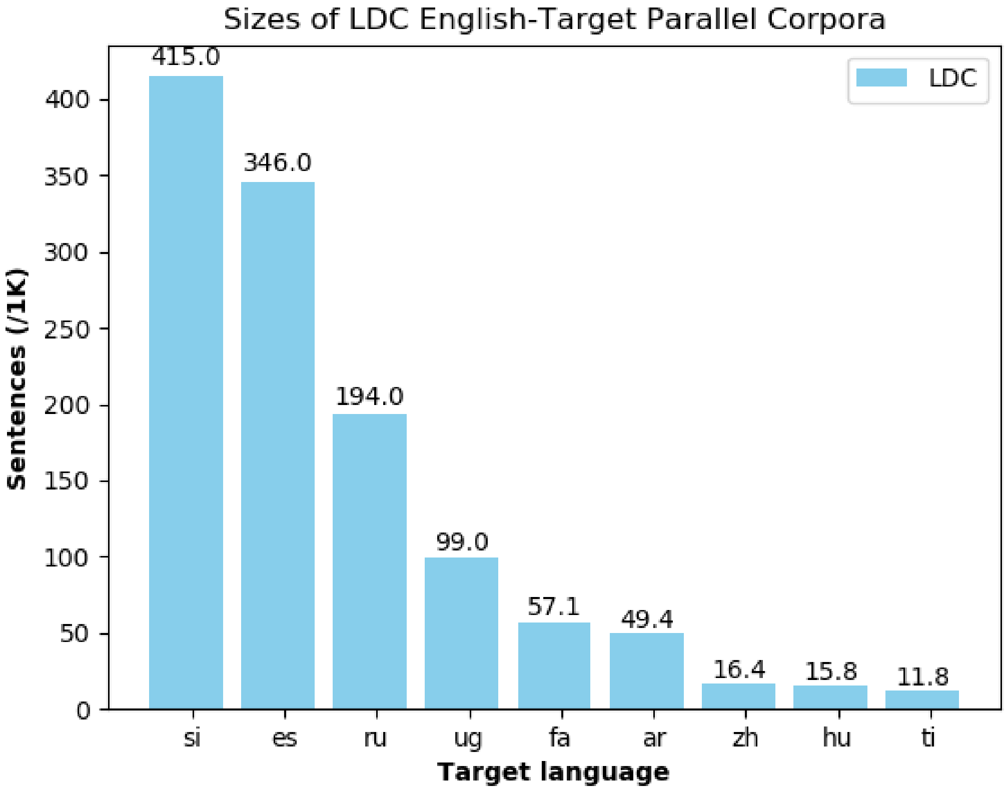


Figure 3.2: Sizes of English-to-target LDC parallel data.

The Linguistic Data Consortium (LDC) parallel data, provided under the LORELEI program, is such a corpus. The parallel corpora are included in the LORELEI representative and incident language packages and consist of a combination of the following genres: news, discussion forums, and Twitter. Since a large part of the content concerns political and social matters, including tweets, it is the closest in domain to most of our evaluation data and to the task of sentiment analysis.

The LDC parallel data contains translations from English to nine target languages: Arabic[6], Tigrinya[7], Sinhalese, Persian[8], Spanish[9], Russian, Chinese[10], Uyghur[11], and Hungarian[12]. The corpora are variable in size for the different languages, ranging from only 11.8K sentences (Tigrinya) to 415K sentences (Sinhalese). While the LDC corpus (except in the case of Sinhalese) is smaller compared to the other parallel corpora, it is closer in domain to the data that arises during sentiment analysis.

Figure 3.2 shows the sizes of the English-to-target LDC parallel data.

3.4.2 European Parliament

The EuroParl (EP) corpus (Koehn, 2005) consists of translations of the proceedings of the European Parliament. It has traditionally been used as a machine translation corpus, but less so for cross-lingual sentiment analysis. While it is considered out-of-genre compared to most of our evaluation data, it is closer in domain than religious corpora because it consists of contemporary political text, which is more topically similar to our sentiment evaluation data compared to religious text. EuroParl data is available for 10 of our European languages: Bulgarian, German, English, Spanish, Hungarian, Polish, Portuguese, Slovak, Slovene, and Swedish. (It excludes Russian and Croatian.) The corpus is multi-parallel; i.e, it comprises the same texts translated

[6] LDC2016E89_LORELEI_Arabic

[7] LDC2017E27_LORELEI_IL5_Incident_Language_Pack_for_Year_2_Eval

[8] LDC2016E93_LORELEI_Farsi

[9] LDC2016E97_LORELEI_Spanish

[10] LDC2016E30_LORELEI_Mandarin

[11] LDC2016E57_LORELEI_IL3_Incident_Language_Pack_for_Year_1_Eval

[12] REFLEX_Hungarian_LDC2015E82_V1.1

among all languages, and our data consists of 294K sentences for each language.

3.4.3 Bible and Quran

Religious text is an out-of-domain and out-of-genre source of data. It is an unconventional choice of resource for sentiment analysis, as the genre, domain and vocabulary are quite different from the typical sentiment analysis evaluation text. However, such corpora offer several advantages (Christodouloupoulos and Steedman, 2014) because of their size and availability of languages compared to EuroParl.

We have used the Bible corpus of Christodouloupoulos and Steedman (2014), which contains Bible translations for 100 languages, and the Tanzil translations for the Quran[13] to create a combined parallel corpus (QB). The QB corpora are available for sixteen of the eighteen evaluation languages; digital versions of the corpora are not available for Tigrinya or Sinhalese.

The number of available translations varies by languages; moreover, the Bible corpus does not include Uyghur translations but the Quran does. Figure 3.2 shows the sizes of the English-to-target LDC parallel data.

3.5 Comparable Corpora

Unlike parallel corpora, comparable corpora do not contain sentence-aligned translations between the source and target languages, but instead consist of text that is similar in both languages, such as texts describing the same topic or news event. The central advantage of comparable corpora is that they are much larger and more easily available for a greater number of languages. Moreover, comparable corpora can be purposefully selected to be in-domain to topics that we choose. The disadvantage, on the other hand, is that comparable corpora texts do not correspond to direct transla-

[13] http://tanzil.net/trans/

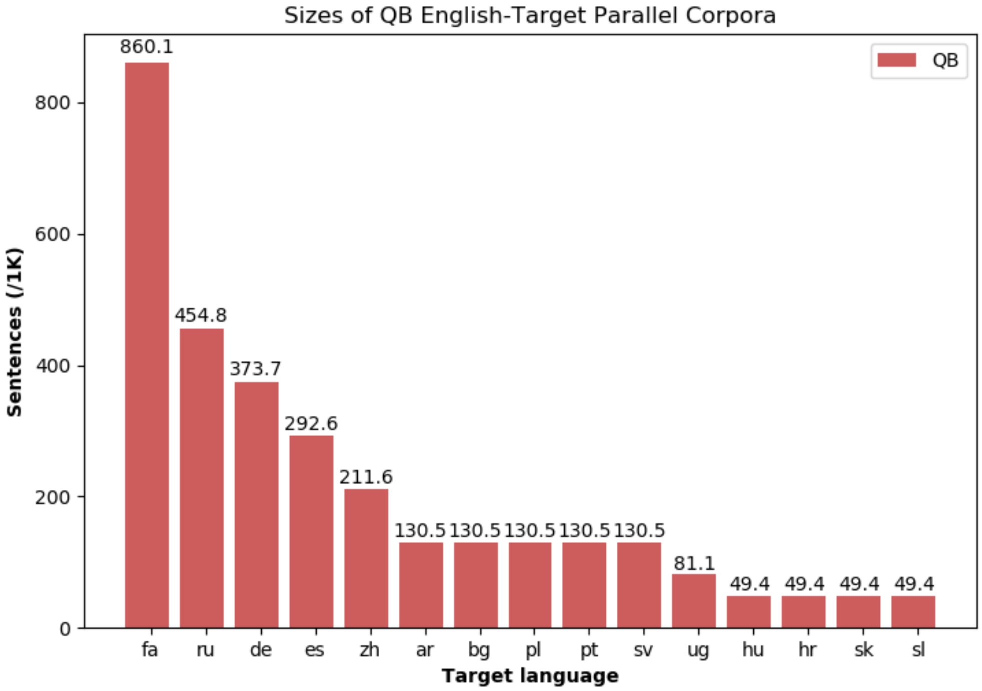


Figure 3.3: Sizes of English-to-target QB parallel data.

tions between source and target languages, making this mode of transfer potentially less accurate than that of parallel corpora.

This is the first work, to the best of our knowledge, that creates and uses comparable corpora for transferring sentiment cross-lingually and that demonstrates the results effectively across a number of languages. Similar to parallel corpora, comparable corpora are used in our models mainly to generate feature representations in a bilingual or interlingual space, which can then be fed into a language-agnostic classification model. In following chapters, we describe how we use comparable corpora from Wikipedia for cross-lingual sentiment transfer, and we measure the extent to which comparable corpora methods suffer from inaccuracy compared to out-of-domain parallel corpora.

3.5.1 Wikipedia

To create a comparable corpus, we collected Wikipedia articles about similar topics in each of the source and target languages. We picked a set of broad pre-defined topics and used the Wikipedia API[14] to query articles about these topics in each of the languages.

We chose 61 broad pre-defined topics, intended to cover a broad range of domains (e.g politics, science, sports), including named political and geopolitical entities relevant to the target languages (e.g Xinjiang, Barack Obama), and translated them from English to the evaluation language either by consulting the Native Informant - translating the keywords takes about 15 min of the Native Informant's time - or by using Google Translate when available. These topic words can be thought of as seed words for creating a cross-lingual linguistic resource. We limited the maximum number of Wikipedia articles retrieved to 1000 per topic word. The topic keywords that were translated into target languages and used for querying the corpus are shown in Table 3.11.

Comparable Corpus Topic Words
politics war terrorism sports entertainment culture environment health economics society education science technology food history mathematics nature geography people art philosophy religion medicine computers law agriculture Obama Trump Clinton Putin ISIS Syria Iraq America Africa Asia China India Europe Arab Germany Spain Hungary Poland Portugal Palestine Israel Iran Pakistan Kazakhstan Xinjiang Bulgaria Slovakia Slovenia Croatia Russia Sweden Rwanda Sri-Lanka Ethiopia Eritrea

Table 3.11: Topic words for querying Wikipedia comparable corpus.

This corpus covers the most languages out of all the presented corpora. We collected it for all eighteen evaluation languages including Uyghur and Sinhalese; however, we exclude the Tigrinya corpus from analysis because of the exceedingly small size of the resultant corpus.

[14] https://pypi.python.org/pypi/wikipedia

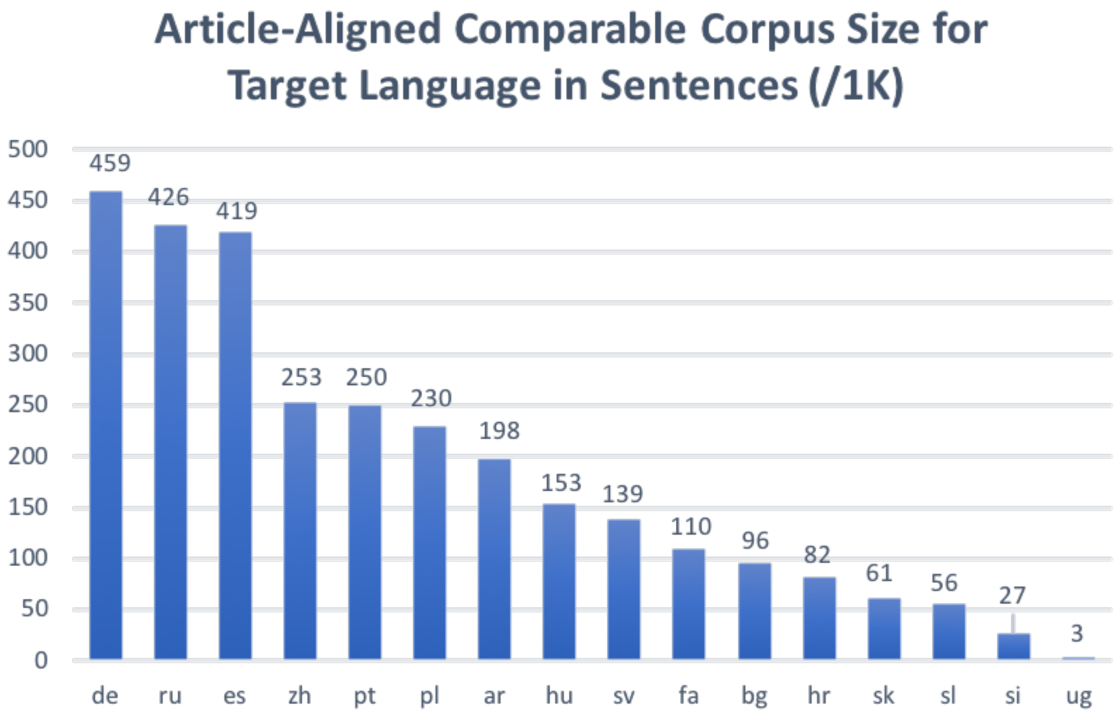


Figure 3.4: Sizes of English-to-target article-aligned comparable data.

3.5.2 Aligning the Comparable Corpora

We queried two versions of the corpus: the first is aligned across all languages by the 61 topics (t_s, t_t), and the second is aligned between English and the target language by articles (d_s, d_t). For the article-aligned corpus *wiki-article*, we align articles only if they correspond to language-linked articles on Wikipedia. This latter corpus results in better cross-lingual sentiment performance and will be referred to as our main comparable corpus throughout the work.

The full sizes of the topic-aligned extracted corpora are 2.1M sentences (English), 1.6M (German), 1.6M (Spanish), 1.6M (Russian), 1.5M (Hungarian), 1.4M (Slovene), 1.2M (Croatian), 1.1M (Polish), 1M (Portuguese), 1M (Arabic), 1M (Chinese), 973K (Slovak), 938K (Bulgarian), 790K (Persian), 668K (Sinhalese), 671K (Swedish), 236K (Uyghur), and 5.3K (Tigrinya). The sizes of the English-to-target article-aligned corpora are shown in Figure 3.4.

	% MPQA	% P	% N	% O
QB	9.0	47.6	32.8	19.5
LDC	8.8	45.5	30.4	23.8
Comparable	5.5	39.2	31.7	28.9
EP	12.4	46.7	19.5	33.6

Table 3.12: Distribution of MPQA tokens vs. total tokens, and among sentiment labels (P:Positive, N:Negative, and O:Neutral) in English side of translation corpora. The English-Arabic corpus was used for LDC, QB, and Comparable; for EP, the corpus is multi-parallel.

3.5.3 Sentiment Content

We assessed the sentiment content of our translation corpora (i.e LDC, EP, QB, and *wiki-article* Comparable), by tagging the English side of the corpus with the MPQA subjectivity lexicon (Wilson et al., 2005) and computing the distribution of subjective positive, negative, and neutral labels among tokens that are identified by the lexicon.

We see that the Quran-Bible corpus has the lowest subjective neutral content (19.5% of lexicon tagged words) and the highest sentiment content (80.4%), while the EuroParl corpus has the highest subjective neutral content (33.6%) and lowest sentiment content (66.2%), especially negative content (19.5%). The article-aligned comparable corpus is most evenly distributed for sentiment, followed by the LDC corpus. In our cross-lingual sentiment experiments, we find that this sentiment content affects the pre-training of our bilingual sentiment features; namely, that EuroParl is less helpful for this purpose while LDC, which has higher sentiment content, is more helpful.

3.6 Monolingual Corpora

Monolingual corpora are the most likely available of corporal resources. While they are more abundant for high-resource languages, they are available for all our evaluation languages. Monolingual corpora are used mostly to generate monolingual and bilingual feature representations to be used by the sentiment analysis models.

We used monolingual data from the Wikipedia language dumps[15] and from the LDC representative and language packs. We used Wikipedia monolingual data for all languages except Tigrinya and Sinhalese, where we instead relied on the monolingual LDC data provided as part of the incident language packs[16][17].

Table 3.13 shows a summary of language resources by corpus.

Language	LDC	EP	QB	wiki-article	Mono
Arabic	X	–	X	X	X
Tigrinya	X	–	–	–	X
Sinhalese	X	–	–	X	X
Persian	X	–	X	X	X
English	X	X	X	X	X
German	–	X	X	X	X
Swedish	–	X	X	X	X
Spanish	X	X	X	X	X
Portuguese	–	X	X	X	X
Bulgarian	–	X	X	X	X
Croatian	–	–	X	X	X
Polish	–	X	X	X	X
Russian	X	–	X	X	X
Slovak	–	X	X	X	X
Slovene	–	X	X	X	X
Mandarin Chinese	X	–	X	X	X
Uyghur	X	–	X	X	X
Hungarian	X	X	X	X	X

Table 3.13: Summary of language resources by corpus.

[15] https://dumps.wikimedia.org/arwiki/latest/

[16] LDC2017E27_LORELELEI_IL5_Incident_Language_Pack_for_Year_2_Eval_V1.1

[17] LDC2018E57_LORELELEI_IL10_Incident_Language_Pack_for_Year_3_Eval_

3.7 Measuring Comparability

The comparability of a bilingual corpus reflects the degree of similarity between between the texts in the source and target languages, and therefore provides a means of measuring the potential of a resource for being an effective medium for cross-lingual transfer. We describe two measures for assessing the comparability of our bilingual resources. The first is the translation comparability measure introduced by Li and Gaussier (2010), and the second is an extension we propose for measuring the sentiment comparability of a bilingual corpus.

3.7.1 Translation Comparability

Li and Gaussier (2010) define the comparability M of a bilingual corpus C_e, C_f with vocabularies ${C_e}^v, {C_f}^v$ as the expectation of finding a dictionary translation for a source word $w_e \in {C_e}^v$ in the target vocabulary ${C_f}^v$, or for a target word $w_f \in {C_f}^v$ in the source vocabulary ${C_e}^v$. The measure is computed by finding the proportion of words in a bilingual dictionary that get translated in the bilingual corpus, and has been shown to be correlated with gold-standard annotated comparability assessments. Given the source corpus C_e, the target corpus C_f, and a bilingual dictionary D_{ef}, the following symmetric measures are computed:

$$M_{ef} = \frac{1}{|{C_e}^v \cap D_e^v|} \sum_{w_e \in {C_e}^v \cap D_e^v} \sigma(w_e, {C_f}^v) \tag{3.1}$$

$$M_{fe} = \frac{1}{|{C_f}^v \cap D_f^v|} \sum_{w_f \in {C_f}^v \cap D_f^v} \sigma(w_f, {C_e}^v), \tag{3.2}$$

where $\sigma(w_e, {C_f}^v)$ is 1 if w_e has a translation in the target corpus and 0 otherwise, and similarly $\sigma(w_f, {C_e}^v)$ is 1 if w_f has a translation in the source corpus and 0 otherwise.

M_{ef} measures the proportion of source words in the dictionary that have a translation in the target corpus, and M_{fe} measures the proportion of target words in the dictionary that have a translation in the source corpus. The comparability measure M is then computed as follows:

$$M = \frac{\sum_{w_e \in {C_e}^v \cap D_e^v} \sigma(w_e, {C_f}^v) + \sum_{w_f \in {C_f}^v \cap D_f^v} \sigma(w_f, {C_e}^v)}{|{C_e}^v \cap D_e^v| + |{C_f}^v \cap D_f^v|}. \tag{3.3}$$

It thus computes the proportion of all dictionary words, source and target, that have a translation in the corresponding bilingual corpus.

3.7.2 Sentiment Comparability

To compute the sentiment comparability of a bilingual corpus, we propose a simple extension: we measure the proportion of source language words that are in both the bilingual dictionary *and* in a *source* language sentiment lexicon L_e that get translated in the target corpus.

$$S = \frac{1}{|{C_e}^v \cap D_e^v \cap L_e|} \sum_{w_e \in {C_e}^v \cap D_e^v \cap L_e} \sigma(w_e, {C_f}^v) \tag{3.4}$$

This measure gives us an idea of how much of the sentiment content in the source corpus is translated to the target corpus. The assumption is that a sentiment lexicon does not exist in the target low resource language, and therefore there is no symmetric computation using a target language lexicon.

Additionally, we may also measure the proportion of positive, negative, or neutral words that get translated in order to gain a better idea of the different sentiment content in the source and target corpus. For example, if more negative words have translations than positive words in a bilingual English-Tigrinya target corpus, and thus higher negative sentiment comparability, this would indicate that the Tigrinya corpus has a relatively higher negative sentiment content compared to the English corpus.

$$S_{pos} = \frac{1}{|{C_e}^v \cap D_e^v \cap {L_e}^{positive}|} \sum_{w_e \in {C_e}^v \cap D_e^v \cap L_e^{positive}} \sigma(w_e, {C_f}^v) \tag{3.5}$$

$$S_{neg} = \frac{1}{|{C_e}^v \cap D_e^v \cap {L_e}^{negative}|} \sum_{w_e \in {C_e}^v \cap D_e^v \cap L_e^{negative}} \sigma(w_e, {C_f}^v) \tag{3.6}$$

We keep these metrics in mind as we study the errors made by our cross-lingual models. For instance, ${S_{neg}}^{LDC}$ for Tigrinya is higher than ${S_{pos}}^{LDC}$, and this is also reflected in the F-measure performance of predicting negative sentiment classes in Tigrinya, which we find to be higher than that of predicting positive sentiment classes.

Figure 3.5 shows the translation comparability and sentiment comparability of the LDC parallel corpora, the EuroParl parallel corpora, the QB parallel corpora, and the article-aligned comparable corpora we collected. For D_{ef}, we used the bilingual dictionaries of Rolston and Kirchhoff (2016). For L_e, we used the MPQA English subjectivity lexicon (Wilson et al., 2005).

There are several noteworthy observations to be made. First, translation comparability of EP and LDC corpora are highest, indicating that both have strong potential to be used as a resource for a cross-lingual transfer task. Second, and interestingly, translation comparability of QB is not higher, and is in fact often lower than that of the comparable corpora *wiki-article*. This indicates that either the Wikipedia corpus in fact contains many more exact translations between the source and target languages than we expect for a comparable corpus, or that the source and target vocabularies of Wikipedia articles are more similar to each other than those that exist in Quran and Bible translations, even if they are not direct translations. Third, comparability is higher for languages with larger target corpora sizes and greater similarities to English: EuroParl languages having the highest comparability and sentiment comparability (0.8-1.0), Arabic and Persian having moderate comparability and sentiment comparability (0.6-0.8), and Uyghur, Sinhalese, and Tigrinya

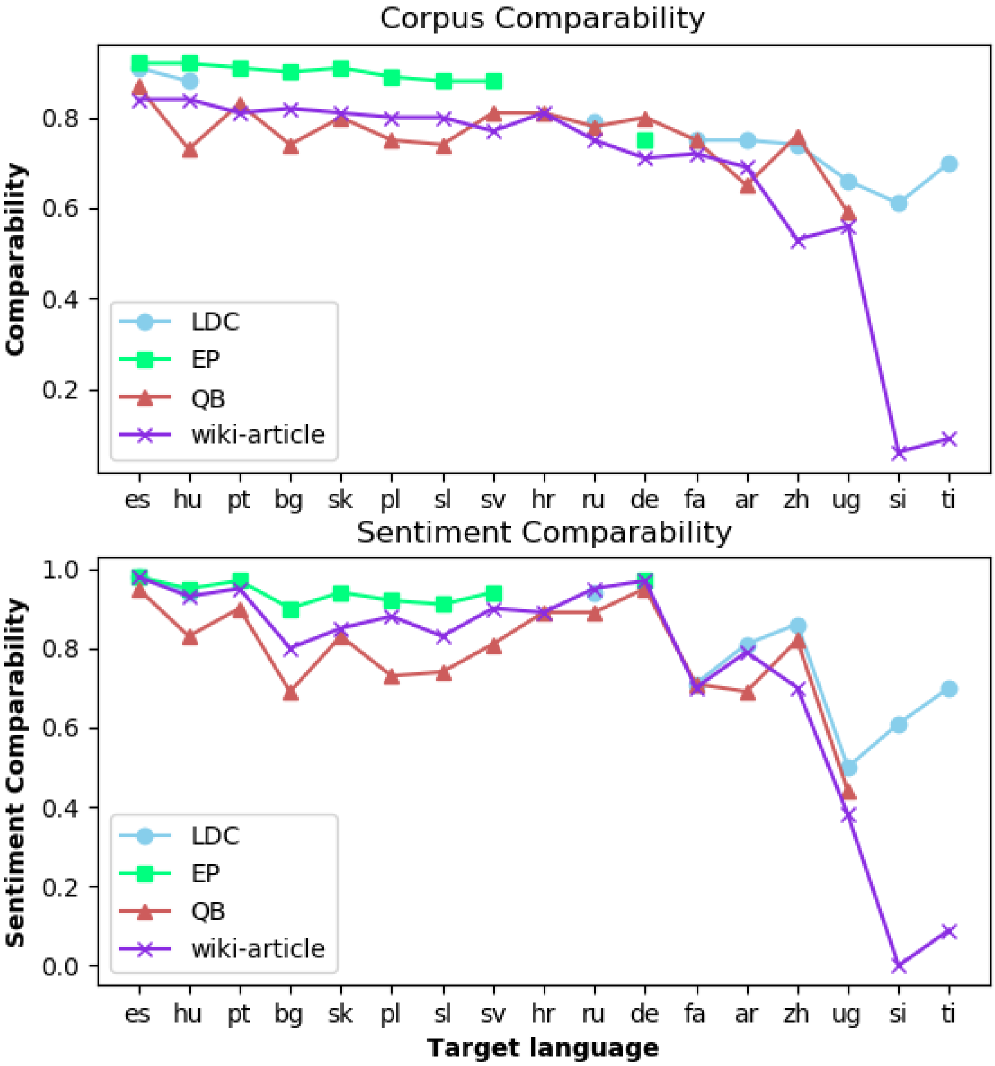


Figure 3.5: Corpus comparability and sentiment comparability for English-target corpora.

having the comparability at the lowest end of the range.

Finally, sentiment comparability is higher than translation comparability for many target languages, indicating that it is easier to transfer sentiment content across the bilingual corpora than it is to transfer all translation content. With sentiment comparability, the difference between *wiki-article* and QB is even more pronounced. However, for the article aligned comparable corpora of Sinhalese, Tigrinya and Uyghur, the *wiki-article* sentiment comparability is quite low, due to the relatively smaller size of their Wikipedia corpora and therefore a small number of words that coincide in the MPQA lexicon, the bilingual dictionary, and the corpus. This also suggests that transferring cross-lingual sentiment will be more difficult for our low-resource languages.

3.8 Conclusion

We collected and made resources available for the study of cross-lingual sentiment analysis and described them in detail; the resources created include three sentiment datasets for Arabic, four native-annotated sentiment evaluation sets for non-Indo-European target languages, and a comparable corpus of topic-aligned and document-aligned Wikipedia articles for 18 languages. We studied the sentiment content and the comparability, including sentiment comparability, of our parallel and comparable corpora and found that in-genre (*LDC*) and in-domain (*LDC*, *EP*, *wiki-article*) translation corpora have higher sentiment comparability than out-of-domain corpora (*QB*); however, all have sufficiently high corpus comparability and sentiment comparability for use in a cross-lingual sentiment analysis task. In the next chapter, we describe approaches for transferring sentiment cross-lingually while making full use of available resources.

Chapter 4

Transferring Sentiment Cross-lingually

Once bilingual resources are identified for bridging the source and target language, they can be used to facilitate the development of cross-lingual models that need only be trained on the source language dataset. In doing so, we can transfer sentiment from a high-resource source language to a low-resource target language, without relying on a machine translation system.

The main challenge with cross-lingual transfer is that most common features from the training dataset do not generalize beyond the source language: for example, lexical features in one language are unlikely to appear in other languages. Our transfer models take this into account by employing a number of techniques for representing features bilingually: through pre-trained cross-lingual word embedding vectors, bilingual sentiment embedding vectors, and lexicalization of the target language to allow for updatable bilingual embedding weights.

The success of cross-lingual transfer models, however, also depends to a large extent on the resources used for building the bilingual feature representations: whether they come from in-genre and in-domain parallel corpora more relevant to the sentiment classification task, out-of-genre parallel corpora such as the Bible and Quran resources, or comparable and monolingual corpora. The availability of these resources varies depending on the target language, necessitating that the different resources be evaluated and that the model be capable of leveraging all of them. Furthermore, several recent works (Gouws and Søgaard, 2015; Conneau et al., 2017; Artetxe et al., 2018) exist that have been able to leverage purely monolingual corpora, sometimes with a bilingual dictionary, to build cross-lingual feature representations for natural language processing. If successful, these methods would be highly advantageous for transferring sentiment; however, there is no evaluation as of yet as to how monolingual-based methods actually perform on a cross-lingual sentiment task with a low-resource language. Our extensive experimental analysis assesses these methods under supervised and unsupervised conditions, and demonstrates that access to a relatively small, well-chosen translation corpus is often preferable when it is available.

One other less explored avenue of potential value is that of transferring not only lexical content, but sentiment information and context itself bilingually through pre-trained cross-lingual embeddings. Utilizing knowledge of sentiment context promises to better estimate the sentiment of target language words than simply translating or projecting words; for example, if the word ‘rescued’ appears in more positive contexts in a pre-training corpus, it will contribute more positive content in a cross-lingual model. In order to take advantage of the sentiment content in translation corpora, therefore, we also present a method for pre-training bilingual sentiment embeddings and bilingual sentiment weights, using only a source-language sentiment lexicon. Furthermore, the sentiment embeddings and weights can be updated during training, if a bilingual dictionary is available. The sentiment weights can be integrated in the cross-lingual model as an additional bilingual representation feature and together with the sentiment embeddings they improve the performance of the model for a majority of languages and under a number of resources.

This chapter describes all the techniques and experimental analyses mentioned above for developing cross-lingual models that transfer sentiment from a high-resource source language to a low-resource target language, with English as a source language. It combines our work on pre-trained cross-lingual embeddings for sentiment analysis (Farra and McKeown, 2019) and our collaborative work on cross-lingual sentiment transfer models (Rasooli et al., 2018). Section 4.1 describes the architecture of the

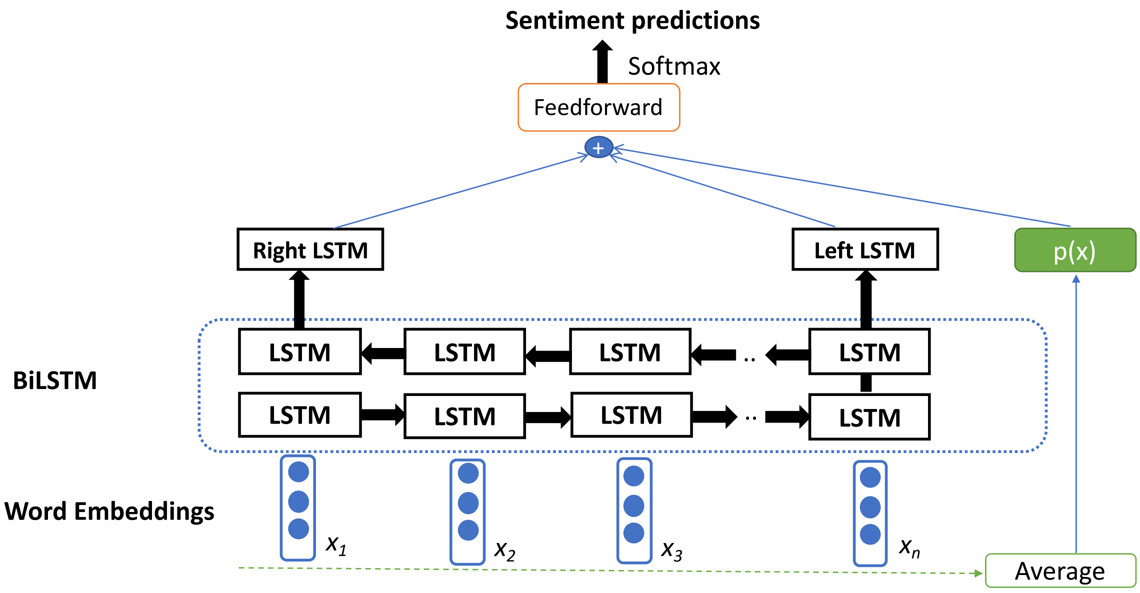


Figure 4.1: Transfer model architecture.

cross-lingual model. Section 4.2 describes our methods for representing features bilingually: pre-trained cross-lingual word embeddings, bilingual sentiment embeddings, and lexicalization of the target language. Section 4.3 dives deeper into the cross-lingual word embeddings, discussing how we create word embeddings vectors using each of the different resources, including comparable corpora. Sections 4.4 and 4.5 present our experiments and results where we evaluate our methods, present extensive experimental analyses, and compare with previous work. We analyze model errors in section 4.6 and conclude in section 4.7. Our code is publicly available[1].

4.1 Transfer Model Architecture

Our base model uses a deep learning architecture with long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). The model accepts a sequence of words $x = \{x_1, x_2, \cdots, x_n\}$ as input, where n is the length of the sequence. The input is then fed into two separate layers: a bidirectional LSTM $r(x)$, which en-

[1] https://github.com/narnoura/cross-lingual-sentiment

codes a recurrent sequence-based representation of the input, and an average pooling layer $p(x)$, which averages features over all input words. The biLSTM captures the incoming sequence of words, while the averaging layer is meant to address scenarios where the source and target languages have different word orders and structures. The two layers $r(x)$ and $p(x)$ are concatenated before being fed into a final feedforward layer with a *softmax* activation: the softmax function computes probabilities for each of the three output target classes $l \in L = \{positive, negative, neutral\}$ and the class with highest probability is predicted as the sentiment label y.

This base model (Figure 4.1) is inspired from our collaborative work (Rasooli et al., 2018). Additional variations of the base architecture are possible depending on the bilingual feature representations and resources that are chosen for representing the input sentence. We describe these representations in the next sections. In Table 4.1, we present a summary of all cross-lingual model variations to be evaluated in this chapter, which will be introduced in the sections that follow. In sections 4.4 and 4.5, we evaluate the performance of the best cross-lingual model for each target language, followed by separate evaluations of each these features and resources.

4.2 Bilingual Feature Representations

Here we describe the feature representations that allow the cross-lingual model to operate bilingually. They are based on various techniques for allowing words and sentiment to be represented in a bilingual vector space *shared* by the source and target language.

4.2.1 Pre-trained Word Embeddings

The cross-lingual model relies substantially on pre-trained bilingual word embedding vectors. A fixed word embedding layer $x_c \in R^{d_c}$ with parameters set to pre-trained

Bilingual Features	**Acronym**
Cross-lingual Word Embeddings	CW
Target Language Lexicalization	+Lex
Bilingual Sentiment Embeddings and Weights	BSW
Cross-lingual Cluster Embeddings	CL
Sentiwordnet Scores	SWN
Bilingual Resources	**Acronym**
In-domain and In-genre Parallel Corpus	LDC
In-domain Parallel Corpus	EP
Out-of-domain Parallel Corpus	QB
Comparable Article-Aligned Corpus	Comparable
Comparable Topic-Aligned Corpus	Wiki-Topic
Monolingual Corpus	Monolingual
Embedding Generation	**Acronym**
Dictionary Code-Switch	Dict-CS
Monolingual Mapping 1	VecMap
Monolingual Mapping 2	MUSE
Bilingual-based	BL

Table 4.1: Summary of cross-lingual model variations. Resources referred to are LDC: Linguistic Data Consortium, EP: European Parliament, QB: Quran and Bible, along with monolingual and comparable corpora described in Chapter 3.

weights, is therefore included in the cross-lingual model. The bilingual word vectors $v_c \in R^{d_c}$ for words $w \in V_{source+target}$, where V is the combined vocabulary of the two languages, are trained differently depending on the translation resources available for the target language - parallel, comparable, or monolingual. We describe these in detail in Section 4.3.

4.2.2 Bilingual Sentiment Embeddings and Weights

As an alternative to bilingual embeddings pre-trained on only lexical context, the cross-lingual model accepts bilingual sentiment embeddings pre-trained on combined

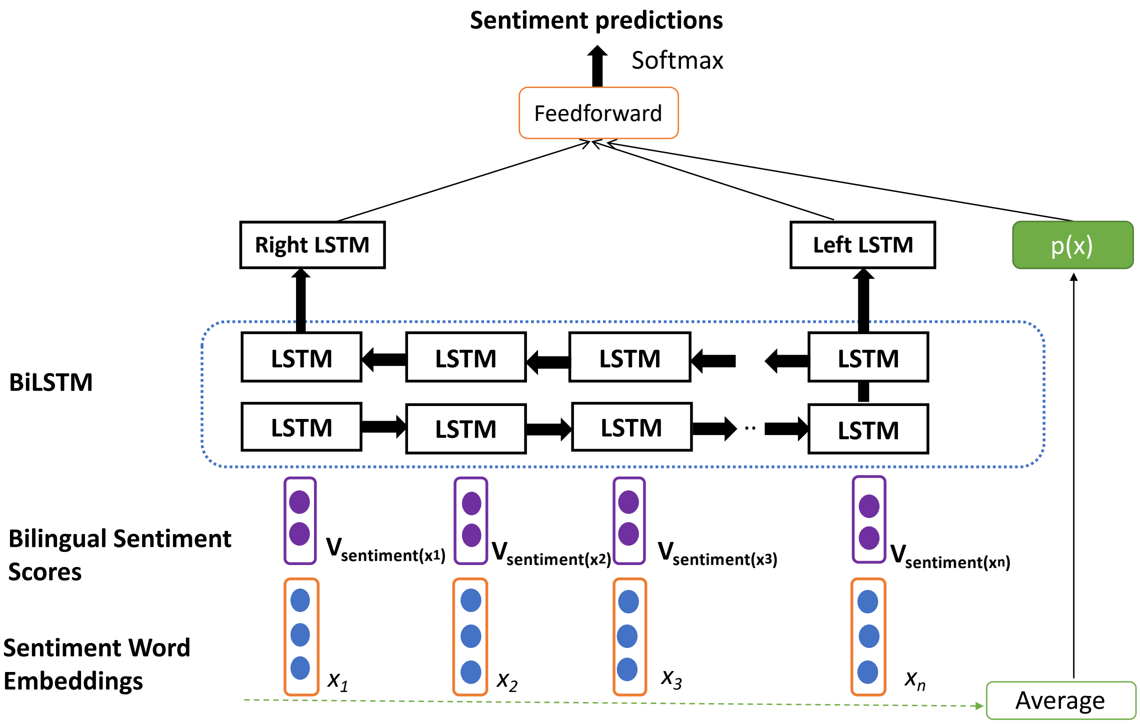


Figure 4.2: Transfer model architecture with bilingual sentiment embeddings and scores.

lexical *and* sentiment context. Our approach for pre-training sentiment embeddings on a translation corpus yields bilingual sentiment embeddings as well as bilingual sentiment output weights, which can be used to create bilingual sentiment scores for each word. The sentiment embeddings replace the fixed embeddings $v_c \in R^{d_c}$ for each word $w \in V_{source+target}$, and the sentiment weights $W_s \in R^{|L| \times d}$ consist of embedding vectors $v_s \in R^d$ for each sentiment label, $s \in L = \{positive, negative, neutral\}$. These are integrated in the cross-lingual model in a combined word embedding and sentiment sequence as described below. For each word in the input sequence, we compute the vector:

$$v_{sentiment(w)} = v_{c(w)}.{W_s}^T \tag{4.1}$$

where $v_{sentiment} \in R^{|L|}$ and $v_{c(w)}$ is the sentiment embedding vector $\in R^{d_c}$. This essentially computes sentiment scores for each label in L for each word, measuring the likelihood of the word being associated with that sentiment label. In practice, we found that normalizing $v_{sentiment}$ by computing the cosine similarity of v_c and ${W_s}^T$ works well, and we have used this configuration in experiments.

Each word $x_i, i = 1...n$ in the input sequence is then represented by the concatenated vector:

$$v_{in} = v_{c(x_i)} \oplus v_{sentiment(x_i)} \tag{4.2}$$

where $v_{in} \in R^{(d_c+|L|)}$. Unlike one-hot sentiment embeddings, this input sentiment sequence does not impose a hard sentiment label on words; instead, words are modeled bilingually according to their likelihood of being associated with different sentiment labels.

A visualization of the cross-lingual model with the bilingual sentiment features is shown in Figure 4.2.

4.2.3 Target Language Lexicalization and Bilingual Embedding Update

Using the representations described above, our cross-lingual model operates fully on bilingual features without any translation into the target language. However, it also provides the option of target language *lexicalization*, whereby if source language words are found in a bilingual dictionary, they are translated during training into the target language. Since not all words in the source training data will have entries in the dictionary, this results in a *partial translation* of the training data, and thus, a code-switched training corpus. Moreover, the word order of the source language sentence is maintained (i.e, only surface translation occurs).

By including target language words in the training, lexicalization allows the distri-

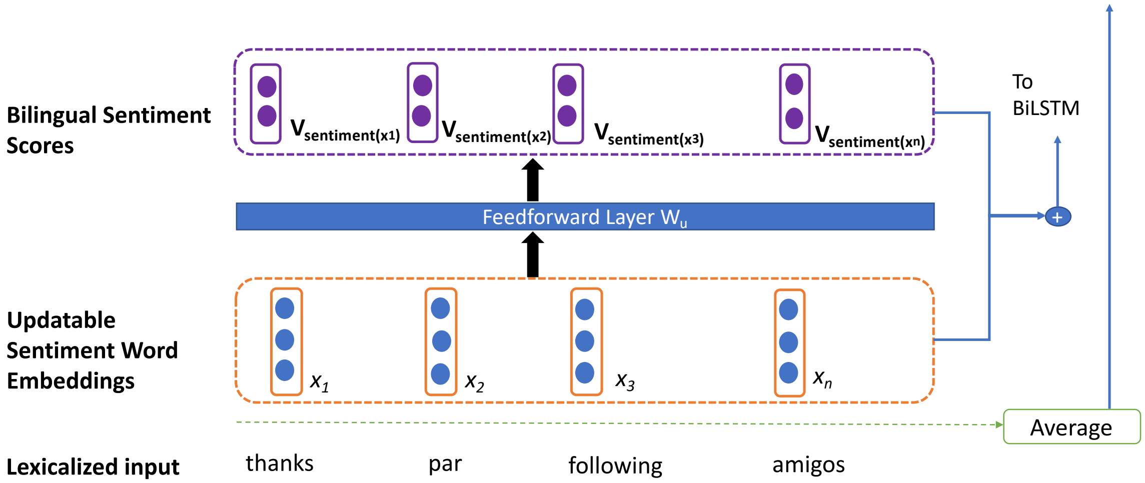


Figure 4.3: Transfer model architecture with updatable bilingual sentiment embeddings and weights, and lexicalized input. The English tweet *'thanks for following friends'* is partially lexicalized to Spanish.

bution of the training data to become closer to that of the evaluation data. Moreover, it allows us to update or fine-tune the bilingual embedding weights during training. Thus, in this configuration, we initialize embeddings to pre-trained weights and update them during training.

Target language lexicalization is inspired from the work of Rasooli and Collins (2017) as well as our collaborative work (Rasooli et al., 2018).

4.2.3.1 Updating Sentiment Weights

The bilingual sentiment weights are trained according to the likelihood of associating words with sentiment scores based on a pre-trained corpus. However, these scores may also be updated based on the training data, which has gold sentiment labels. To create updatable bilingual sentiment weights, we pass the sentiment embeddings $v_{c(w)}$ through an updatable feedforward layer f and initialize its weights to W_s:

$$v_{sentiment} = f(v_{c(w)}.{W_u}^T + b), \tag{4.3}$$

where $b = 0$, f is the *relu* activation function (Nair and Hinton, 2010) and W_u is initialized to W_s.

Figure 4.3 shows how we update bilingual sentiment embeddings and weights using an example of training on English and testing on Spanish.

4.2.4 Cluster Embeddings and SentiwordNet

Two additional bilingual representations we consider are cluster embeddings and sentiment scores from the lexicon Sentiwordnet (Baccianella et al., 2010). Both are used in preliminary versions of our model, which used monolingual-based embeddings and was published in our paper (Rasooli et al., 2018). With our best model, which uses bilingual-corpus-based embeddings, we found that cluster embeddings and Sentiwordnet were less impactful on average. We include experiments to this effect, and we additionally use Sentiwordnet scores as a baseline to compare with our bilingual sentiment features.

If lexicalization is applied and the translation for a source word is found during training, the feature (e.g word vector, cluster or Sentiwordnet score) for the target word is used; otherwise, the English feature is used.

4.2.4.1 Cluster Embeddings

We apply Brown clustering, created using the method of Stratos et al. (2014), to word embedding vectors to create cross-lingual cluster embeddings. We then create an additional input channel $x_b \in R^{d_b}$ which is concatenated with the input word embedding channel and which represents each word by its cross-lingual cluster rather than by its individual word vector.

4.2.4.2 Sentiwordnet

Sentiwordnet uses a number of manual seeding and automated extraction techniques sentiment polarity scores for each English word. Only when a bilingual dictionary is available (i.e, in the 'lexicalization' configuration), we translate the Sentiwordnet lexicon to the target language using the dictionary. We use two-dimensional scores $x_{sw} \in R^2$ as an additional concatenated channel that represents the likelihood of a word being positive or negative, similar to the bilingual sentiment scores $v_{sentiment}$. However, unlike $v_{sentiment}$, Sentiwordnet scores rely on directly translating source language words and can only be used when a bilingual dictionary is available. Moreover, the Sentiwordnet scores extracted from the test data are limited by the target language words that are found in both Sentiwordnet and the bilingual dictionary, while in the case of the bilingual sentiment weights, every test word with a pretrained embedding will get a score.

4.3 Creating Bilingual Features with Different Resources

Here we describe how to create pre-trained bilingual embedding features using each of the different resources that may be available to the target language: monolingual corpora, sentence-aligned parallel corpora, and comparable corpora.

4.3.1 Monolingual-based Embeddings

Our approach for pre-training bilingual embeddings from monolingual corpora relies on the use of a bilingual dictionary. First, a dictionary is learned using word alignments generated from a smaller parallel corpus. Then the dictionary is used to translate words at random in monolingual data concatenated from the source and target languages. This creates a 'code-switched' bilingual corpus, on which a monolingual word embedding model, namely that of Mikolov et al. (2013) can be applied directly. To build these embeddings, we use the monolingual corpora described in Chapter 3 and the parallel corpora to generate word alignments. We refer to this approach as 'Dictionary-Code-Switch', or **Dict-CS**.

This method is inspired from the code-switching approach of Rasooli and Collins (2017) and is also similar to the approach of Gouws and Søgaard (2015).

In Sections 4.4 and 4.5, we also evaluate our models in comparison with the monolingual-based embedding representations of Artetxe et al. (2018) (VecMap) and Conneau et al. (2017) (muse).

4.3.2 Bilingual Embeddings from a Parallel Corpus

Instead of using a parallel corpus to create a dictionary - provided that a well-chosen corpus can be made available for the target language - the pre-trained embeddings can be learned directly on the parallel corpus itself, without any additional monolingual data. This allows the bilingual word vectors to directly make use of bilingual translation context, and if sentiment embeddings are incorporated, of bilingual sentiment context.

4.3.2.1 Parallel Corpus Embeddings

To train bilingual embeddings from a parallel corpus, we follow Luong et al. (2015)'s approach, which uses a joint objective of monolingual and bilingual models:

$$\alpha(Mono_1 + Mono_2) + \beta B_i \tag{4.4}$$

Words are first aligned in the source and target language sentences; note that word alignments need not be learned in order to achieve effective bilingual quality, and assuming monotonic alignments - i.e, each source word is mapped to a target language word in keeping with word order - works effectively in practice, as shown by Luong et al. (2015). For each source or target word, both monolingual and bilingual contexts are used to predict the word using the CBOW or skipgram (Mikolov et al., 2013) objective. The joint objective can be thought of as essentially learning four joint 'word2vec' models *src*→*src*, *trg*→*src*, *src*→*trg*, *trg*→*trg* for source and target languages. Figure 4.4 represents a schematic of the four monolingual and bilingual contexts used in this bilingual embedding approach.

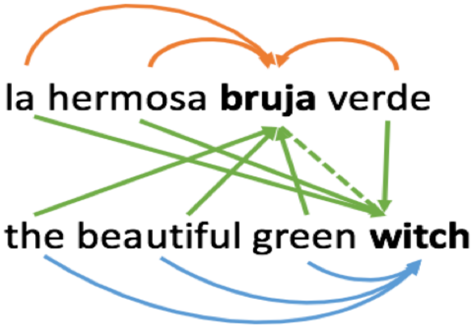


Figure 4.4: Monolingual and bilingual English and Spanish contexts used for predicting 'witch' and 'bruja', which are aligned. The glosses for the Spanish text are: 'the beautiful witch green'.

4.3.2.2 Bilingual Sentiment Embeddings

We extend parallel corpus embeddings by modeling the probability,

$$p(s[w] = s_j | context) \tag{4.5}$$

that a *source-language* word has a prior sentiment label, given the *source* or *target* words in its translation context. Thus, the CBOW objective is extended as follows. We assume our corpus consists of pairs of words and contexts $D = (w, context)$ and pairs of sentiment labels and contexts $S = (s, context)$, $s_j \in L = \{positive, negative, neutral\}$. Words and contexts can belong to source *src* or target *trg* languages, while sentiment labels belong only to source *src* languages.

For source words that have lexicon entries, we model the monolingual and bilingual CBOW and sentiment objectives:

$$\underset{\theta}{\operatorname{argmax}} \sum_{(w,c)\in D} log \frac{1}{1+e^{-v_{context}.v_w}} + \gamma log \frac{1}{1+e^{-v_{context}.v_s[w]}} \tag{4.6}$$

where (w, c) includes both source and target language contexts, $s[w]$ is the sentiment label assigned to source word w, and γ is a hyperparameter indicating the effect of the contribution of sentiment labels.

For target words, and source words without lexicon entries, whose prior sentiment labels are assumed unknown, we model the monolingual and bilingual CBOW objectives,

$$\underset{\theta}{\operatorname{argmax}} \sum_{(w,c)\in D} log \frac{1}{1+e^{-v_{context}.v_w}} \tag{4.7}$$

where $v_w \in R^d$ is the word vector representation of w and the context vector,

$$v_{context} = \frac{1}{2b} \sum_{i\in[-b,b]-\{0\}} v_c(w_i) \tag{4.8}$$

is an average of context word vectors $v_c \in R^{d_c}$ in a window $\{-b, b\}$ around the center word w.

The parameters to be learned jointly in bilingual space are $\theta = \{v_w, v_c, v_s\} \in R^{d_c}$ for all w, c in $V_{source+target}$ and all sentiment labels s_j, where v_w and v_c are the output and input word embedding weights, and v_s are the output bilingual sentiment embedding weights. The parameters are learned using gradient descent and negative sampling. Thus, the input weights w_c and the output sentiment weights v_s are updated using an additional sentiment error term.

As a result, we effectively induce a sentiment prior on word embedding vectors, and the model learns the sentiment associated with source and target words in bilingual contexts, whose sentiment label was previously unknown. Source and target words with a similar sentiment label distribution - how likely they will be labeled 'positive' or 'negative' in most contexts - will cluster more closely in the bilingual embedding space. The model is applicable monolingually as well as bilingually, without explicitly modeling any word alignments.

In addition to learning input and context word vectors, the model produces pre-trained bilingual sentiment vectors or bilingual sentiment weights $v_s \in R^{d_c}$, $s \in \{positive, negative, neutral\}$ which can be utilized in the cross-lingual sentiment model to create sentiment scores as described in Section 4.2.2. We have used the MPQA lexicon with 'priorpolarity' sentiment tags (Wilson et al., 2005) as a seed lexicon to tag all source language words in our bilingual embedding corpora.

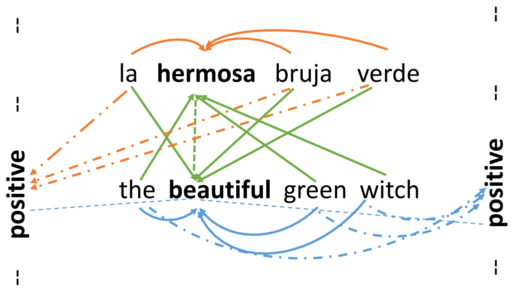


Figure 4.5: Monolingual and bilingual English and Spanish word and sentiment contexts used for predicting 'hermosa', 'beautiful', and output sentiment vector v_s for the 'positive' sentiment label. 'Beautiful' and 'positive' are aligned using an English sentiment lexicon. The remaining three words do not have any entry in the lexicon, and are represented by dashes '- -'.

Figure 4.5 shows a schematic of bilingual word and sentiment contexts used for learning the embeddings. If similar examples occur often enough in our pre-training corpus, the vectors for the words ‘witch’ and ‘bruja’ will become closer to the ‘positive’ vector and will be more likely to be associated with positive sentiment.

4.3.3 Bilingual Embeddings from a Comparable Corpus

We next describe how to leverage comparable corpora to create bilingual word embeddings and sentiment embeddings.

4.3.3.1 Comparable Corpus Embeddings

As described in Chapter 3, comparable corpora can be aligned by documents (d_s, d_t) or more broadly by topics (t_s, t_t); the first option is more precise while the second gives us more data. We found that the article-aligned corpus ‘wiki-article’ results in greatly enhanced performance by the cross-lingual model, and we use this corpus as our main comparable translation corpus. However, at the end of this chapter, we also include results obtained using the topically aligned corpus.

Our approach for creating bilingual embeddings from comparable corpora uses the ‘length-ratio-shuffle’ method of Vulić and Moens (2016) to construct a pseudo-bilingual document. This approach, while preserving the monolingual order, inserts source and target words w_s, w_t into the bilingual document by iteratively appending R source words from d_{s_j} or t_{s_j} followed by one target word from d_{t_j} or t_{t_j}, where $R = \left\lfloor \frac{m_s}{m_t} \right\rfloor$ is the ratio of source word tokens to target word tokens in document or topic j (assuming without loss of generality that $m_s \geqslant m_t$). Remainder source words are appended in monolingual order at the end of the bilingual document. Monolingual embedding training is then applied to the bilingual document to generate the comparable embeddings. We follow this approach for the article-aligned comparable corpus *wiki-article*.

However, if the corpus is not article-aligned, it becomes much less likely that the order of sentences in the source and target languages reflects actual translations.

Thus, for topic-aligned corpora, instead of preserving the order of sentences within original topics t_j, we first compute Term Frequency - Inverse Document Frequency (TF-IDF) for source words w_s and target words w_t. We then separately rank the source and target sentences by their averaged TF-IDF scores. Since the topics are the same in both languages, this makes it more likely that when we merge the documents, the frequently occurring topical words in the source language will align to the corresponding frequently occurring topical words in the target language. Figure 4.6 shows a diagram of Merge-TFIDF.

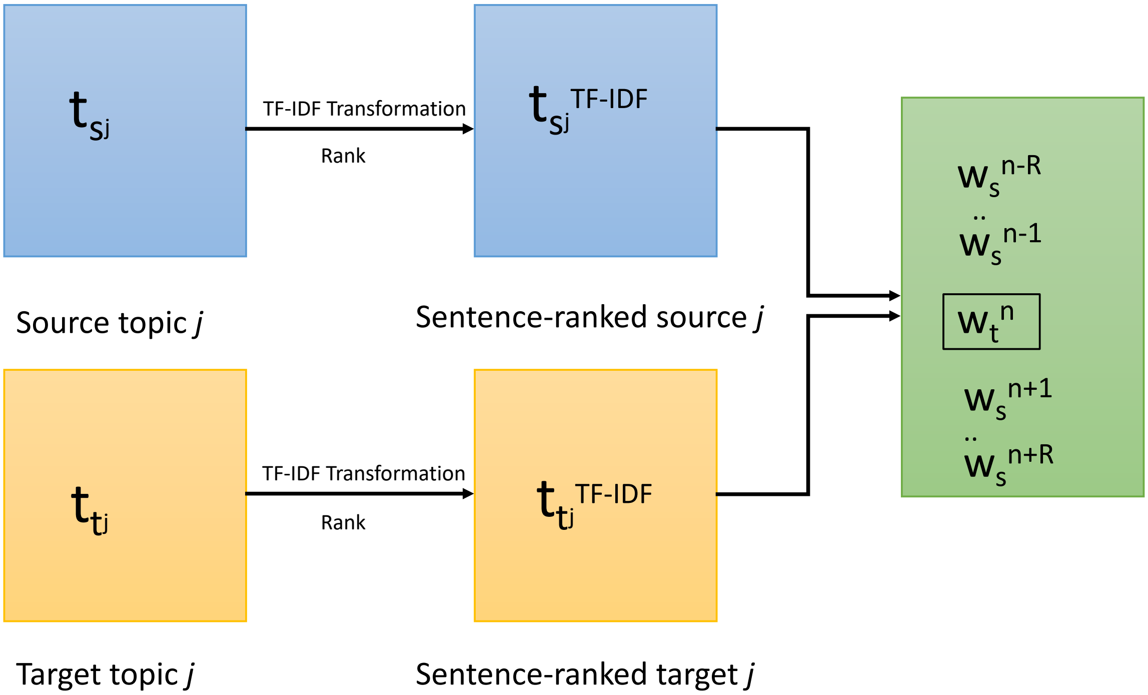


Figure 4.6: Creating a comparable bilingual document with Merge-TFIDF. n is the index of the target language word in the bilingual document.

4.3.3.2 Comparable Corpus Bilingual Sentiment Embeddings

Our approach to bilingual sentiment embeddings can be applied to a non-parallel comparable corpus while specifying a large window size b, as learning sentiment in context is less dependent on word order than are tasks like machine translation.

Thus, producing efficient bilingual sentiment representations of words may be possible without a sentence-aligned corpus. The process for building bilingual sentiment embeddings from a comparable corpus is outlined as follows:

1. We align language-linked articles by documents (d_{src}, d_{trg}).
2. We build code-switched documents using the 'length-ratio-shuffle' method of Vulić and Moens (2016) as described above.
3. We run our sentiment lexicon training objective in the monolingual configuration.

4.4 Experiments

In our experiments, we evaluate the cross-lingual model variations described thus far in this chapter (Table 4.1) and assess the following factors:

1. **Best Transfer Model**. The best cross-lingual model for each target language amongst among our proposed model variations[2], and its performance compared with a supervised model trained on the same language, if training data is available.
2. **Bilingual Resources**. The performance of the cross-lingual model under different resource availablity: in-domain and in-genre parallel corpus (LDC), in-domain parallel corpus (EP), out-of-domain parallel corpus (QB), comparable corpus (Comparable), and monolingual corpus (Monolingual). These are the corpora that were described in Chapter 3, Sections 3.4, 3.5, and 3.6. The article-aligned comparable corpus 'Wiki-Article' is used as our main comparable corpus, while the topic-aligned corpus 'Wiki-Topic' is evaluated separately. In particular, we study:

[2] These exclude comparisons with VecMap and MUSE, which are detailed in further sections.

   - **Embedding Generation**. Whether and under what conditions bilingual embeddings from translation corpora (BL) are preferable to monolingual embedding-based methods (ML), namely those described in Rasooli and Collins (2017) (Dict-CS), Conneau et al. (2017) (muse), and Artetxe et al. (2018) (vecmap).
   - **Resource Comparison**. How well non-parallel resources (comparable and monolingual) fare compared with out-of-domain parallel resources and in-domain parallel resources.

   Thus, our experiments study the performance of the model under varied *supervised* (a parallel corpus or bilingual dictionary is available) and *unsupervised* (only a non-parallel comparable or monolingual corpus is available) scenarios.

3. **Feature Representations.** The performance of the cross-lingual model when using different bilingual feature representations: pre-trained cross-lingual word embeddings (CW), sentiment embeddings and weights (BSW) and lexicalization of the training data with embedding update (+Lex).

4.4.1 Setup and Configurations

4.4.1.1 Model Configurations

The cross-lingual model was developed and tuned on held-out development sets from Arabic (671 sentences), Bulgarian (2999 sentences), English (5832 sentences), and Persian (1000 sentences). No development sets were used for the remaining target languages. We tuned our monolingual-based (Dict-CS) and bilingual-based (BL) models separately. For the monolingual-based model: the tuned model used 7 epochs for training, a hidden layer size of 400, and a batch size of 10K. For the bilingual-based model: the tuned model used 5 epochs for training, a hidden layer size of 100, and a batch size of 32. We trained the models on the *English source dataset* using the Adam optimizer (Kingma and Ba, 2014), with categorical cross-entropy loss, and applied them to the evaluation datasets of each target language.

4.4.1.2 Bilingual Dictionaries

Bilingual dictionaries were used when lexicalizing the training data, and when training monolingual-based embeddings by code-switching (Dict-CS method). In these configurations, we assume the availability of parallel corpora resources, which can be used to automatically create word alignments. We created source-target word alignments with GIZA++ (Och and Ney, 2003) and extracted bilingual dictionaries automatically by assigning dictionary entries to source-target pairs which are aligned most frequently in the parallel corpus.

4.4.1.3 Embedding Configurations

1. **Bilingual-based embeddings**. For training the bilingual-based word and sentiment embeddings, we used the Multivec (Bérard et al., 2016) toolkit, which provides support for monolingual and bilingual embeddings, namely those of Luong et al. (2015), and we extended the toolkit to provide support for our bilingual sentiment embedding training[3]. Training completes in under 3.2 minutes for the largest parallel corpus size of 415K sentences and in 4.8 seconds for the smallest corpus of 11.8K, on CPU. We built 300-dimensional vectors with a context window size of 5, except for comparable corpus embeddings, where we used a window size of 10. We used $\gamma = 1$ for the sentiment embeddings. Bilingual-based embedding models were trained using the parallel and comparable corpora described in Chapter 3.
2. **Monolingual-based embeddings**. For training our monolingual-based code-

[3] https://github.com/narnoura/multivec

switched embeddings (DICT-CS), we used the Word2vec[4] tool with its default configurations. Monolingual-based models were trained using the monolingual corpora described in Chapter 3 for each of our 18 languages.

3. **VecMap and MUSE**. These models use large monolingual corpora and, under supervised scenarios, bilingual dictionaries, to map monolingual spaces into a shared cross-lingual space. Their implementations are publicly available[5][6]. We built 300-dimensional embeddings using both and kept default parameters. In the supervised versions, we used the same dictionaries that we created using our parallel corpora. For MUSE, we do not produce embeddings or results in the unsupervised case, because their available system and iterative refinement procedure requires an external validation dictionary. In the supervised case, we retained MUSE's validation dictionary if it was available for our language; if it was not, we split 2000 words from our parallel corpus dictionaries and used them for validation.

Table 4.2 shows the vocabulary size of target language words resulting from the bilingual embeddings created from each of the corpora. English vocabulary size is 20.5K words for the European Parliament corpus (EP), 918K words for monolingual Wikipedia, and varies by target language for the remaining translation corpora. Rare words (occurring fewer than five times in the corpus) are filtered out during the embedding training process. Thus, embeddings created from a larger corpus - such as the Arabic Quran and Bible corpus - may still have a smaller vocabulary size than a smaller corpus - such as the Arabic LDC corpus - if it contains more rare words. We can also notice that while the EP corpus contains the same number of translations across all nine languages, the resulting vocabulary size differs significantly

[4] https://code.google.com/archive/p/word2vec/

[5] https://github.com/artetxem/vecmap

[6] https://github.com/facebookresearch/MUSE

among these languages; it is higher for the more morphologically complex European languages (Polish, Slovak, Slovene, and Hungarian) which have vocabulary sizes close to or exceeding 50K words, somewhat smaller for Bulgarian(bg), German(de), and Swedish(sv), and smallest for English, Spanish(es) and Portuguese(pt), which are less morphologically complex.

We built cross-lingual Brown clusters from the generated word embedding vectors, specifying a number of clusters equal to 500.

Language	LDC	EP	QB	Comparable	Monolingual
ar	20.5	–	17.3	66.8	152.0
bg	–	37.5	21.3	37.6	161.2
de	–	36.0	23.8	116.5	1279.5
es	58.6	28.6	24.9	88.0	639.3
en	–	20.5	–	–	918
fa	13.6	–	70.0	23.2	171.9
hu	39.7	57.5	18.6	63.8	388.5
hr	–	–	15.3	43.7	237.8
pl	–	49.7	24.8	84.9	581.1
pt	–	29.9	22.4	60.7	380.3
ru	56.1	–	50.3	130.7	543.7
si	35.3	–	–	19.2	40.7
sk	–	47.3	13.7	31.0	192.4
sl	–	43.6	14.2	30.4	190.2
sv	–	33.2	16.2	45.2	305.2
ti	5.4	–	–	–	59.9
ug	18.9	–	18.3	2.9	19.3
zh	6.9	–	27.3	18.3	466.8

Table 4.2: Target language vocabulary size for embeddings created from all corpora. The acronyms for the parallel corpora in the first three columns are Linguistic Data Consortium (LDC), European Parliament (EP), and Quran and Bible (QB). Comparable refers to the article-aligned Wikipedia corpus and Monolingual refers to the monolingual corpora, both described in Chapter 3. Vocabulary size is represented in 1000 word units.

4.4.2 Data

For the experiments in this chapter, we use the data from the untargeted datasets described in Chapter 3, Section 3.3.1: namely, the English European Twitter dataset (Mozetič et al., 2016) for training, and the untargeted sentiment evaluation datasets for the 17 target languages: Arabic (ar), Bulgarian (bg), German(de), Spanish(es), Persian(fa), Hungarian(hu), Croatian(hr), Polish(pl), Portuguese(pt), Russian(ru), Sinhalese(si), Slovak(sk), Slovene(sl), Swedish(sv), Tigrinya(ti), Uyghur(ug), and Chinese(zh).

4.4.2.1 Text Preprocessing

We preprocessed the text by tokenizing the datasets, translation corpora, and monolingual corpora. We used the Stanford Chinese Segmenter (Chang et al., 2008), MADAMIRA Arabic tokenizer and morphological analyzer (Pasha et al., 2014), Hazm Persian tokenizer[7], European tokenizers in the EuroParl package, OpenNLP[8] and the Moses tokenizer[9].

For these English-to-target experiments, we ran MADAMIRA using the Arabic Treebank (ATB) tokenization scheme with the 'BWFORM' option, a heuristic-based method that is more limited than the default regeneration method, but is sufficient for most tokenization needs. In later chapters on targeted sentiment where Arabic is treated as a high-resource language, we use the default regeneration tokenization option.

We cleaned all tweets by removing hashtags, user mentions, and links.

[7] https://github.com/sobhe/hazm

[8] ttps://opennlp.apache.org/

[9] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

4.4.3 Metric

For evaluating sentiment predictions, we used macro-averaged F-Measure, F-Macro, which is the F-Measure averaged over the three sentiment classes giving equal weight to each class, a standard metric for sentiment analysis (Rosenthal et al., 2017, 2015a). F-Macro provides a better indication of performance on predicting positive and negative sentiment labels than accuracy, especially when working with unbalanced datasets with higher neutral content.

4.4.4 Statistical Significance

Statistical significance was computed for F-Macro using the bootstrap significance test (Berg-Kirkpatrick et al., 2012). Because deep learning models are non-deterministic and are run with random seeds, our experiments are repeated for 5 runs and the averaged result is presented[10]. To compute statistical significance on the averaged result, we used the majority vote of the 5 runs (predicting 'neutral' in case of conflict) and used the corresponding output to test for significance.

4.4.5 Baseline Approaches

We incorporate two rule-based and lexical baselines and a third baseline which uses Sentiwordnet features in the model.

1. **Majority:** The performance of a model which always predicts the *neutral* majority baseline of the English training data. This baseline is aimed at demonstrating how well the cross-lingual model can overcome the majority baseline, particularly when it was trained on a language with different sentiment distribution.

[10] Except for DICT-CS results, which are reported directly from our paper (Rasooli et al., 2018).

2. **SWN-rule:** The performance of a rule-based lexical model which translates Sentiwordnet to the target language using the dictionaries extracted from the Bible and Quran parallel data (or LDC data, if the Bible and Quran corpus is not available). It assigns positive (negative) sentiment, if the *sum* of the positive (negative) scores of a sentence is at least 0.1 higher than the *sum* of the negative (positive) scores. Otherwise it assigns the neutral label.[11]
3. **SWN:** The performance of the full cross-lingual model when using translated Sentiwordnet scores as features in place of the bilingually trained sentiment scores. This baseline is used when evaluating bilingual feature representations in the lexicalization configuration. It is aimed at assessing the value of training bilingual sentiment scores as opposed to simply projecting the sentiment scores obtained from the source language.

4.5 Results

In the following sections, we show our results. Section 4.5.1 *(Best Transfer Model)* shows the performance of our best cross-lingual model for each language, compared with that of the supervised in-language model and the baselines. Section 4.5.2 *(Bilingual Resources)* presents our extensive experimental analysis of cross-lingual sentiment performance using word embeddings created from different resources, including monolingual embeddings created using Dict-CS, vecmap, and muse. Section 4.5.3 *(Bilingual Feature Representations)* presents detailed results assessing the effect of lexicalization and bilingual sentiment features. Finally, Section 4.5.4 shows the performance of our cross-lingual model in comparison with the adversarial transfer model of Chen et al. (2018).

[11] The lexical baseline approach used in our paper is more lenient than this baseline. It used *average* sentiment score rather than sum, which we discovered results in a much lower baseline score.

4.5.1 Evaluation of Transfer Model

Table 4.3 shows the performance of predicting sentiment in each of the target languages using English as a source language.

The first two columns show the performance of the 'Majority' and 'SWN-Rule' baselines. The third column, Transfer, shows the result of our best cross-lingual model for the target language, among available resources (Linguistic Data Consortium (LDC), EuroParl (EP), Quran-Bible (QB), and Comparable Wiki-Article (Comp)) with bilingual-based cross-lingual embeddings (BL), monolingual-based Dict-CS embeddings (ML), and feature representations (Bilingual sentiment embeddings and weights, Lexicalization, Sentiwordnet, and Clusters).

The fourth column (Sup) shows the result of our best supervised model for the target language, if a training dataset is available for the language. This supervised model results from running our deep learning model (Section 4.1) with either monolingual Wikipedia embeddings or updatable word embeddings initialized during training. This number, along with the baselines, is also shown for the source language, English (en).

The last three columns descibe the configurations of the best transfer model: whether it uses monolingual-based Dict-CS embeddings or bilingual-based embeddings (R-type), the bilingual feature representations used (F-type), and the bilingual corpus used for pre-training or for creating a bilingual dictionary (Corpus).

4.5.1.1 Discussion

Table 4.5.1 shows that the best model configuration most often results with bilingual-based features (*BL*), an in-domain parallel corpus (*LDC or EP*), and with an added representation feature (*BSW* and/or *+Lex*).

Transfer vs. Baseline.

We observe that the best model 'Transfer' is able to easily overcome the majority base-

	Baselines		**Model**		**Best Configuration**		
	Majority	SWN-R	Transfer	Sup	R-type	F-type	Corpus
ar	18.6	38.9	45.9†	**55.5**	BL	BSW	LDC
bg	22.4	37.9	49.3†	**57.5**	BL	CW, SWN, +Lex	EP
de	23.9	39.1	49.2†	**58.3**	BL	CW	EP
es	19.3	33.5	44.4†	**51.4**	BL	CW, +Lex	LDC
en	21.0	46.2	–	**65.9**	–	–	–
fa	17.9	37.8	53.0†	**71.3**	BL	CW, CL	LDC
hu	16.5	36.0	49.1†	**63.0**	BL	BSW, +Lex	LDC
hr	12.8	31.9	39.7†	**61.9**	BL	BSW	Comp
pl	13.8	37.3	43.9†	**62.7**	BL	BSW	EP
pt	17.3	33.3	42.5†	**53.0**	BL	BSW, +Lex	EP
ru	20.0	36.4	50.2†	**68.9**	BL	BSW, +Lex	LDC
si	16.1	36.3	35.2	–	ML	CW,CL, SWN,+Lex	LDC
sk	11.9	34.0	40.8†	**68.6**	BL	CW	Comp
sl	20.6	38.3	42.3†	**56.0**	BL	BSW, +Lex	EP
sv	16.1	39.2	49.0†	**62.7**	ML	CW, CL, SWN,+Lex	EP
ti	23.1	34.5	40.9	–	BL	BSW, +Lex	LDC
ug	23.6	38.6	45.2	–	BL	CW	LDC
zh	19.7	43.9	**56.3**†	47.0	ML	CW,CL, SWN,+Lex	LDC

Table 4.3: Macro-averaged F-measure for predicting sentiment labels ‘positive’, ‘negative’, and ‘neutral’ for best cross-lingual model ‘Transfer’ compared with neutral Majority baseline, lexical Sentiwordnet baseline SWN-R (SWN-Rule), and supervised model Sup trained on the same language. Best results are shown in **bold**, results where Transfer outperforms baselines are shown in blue, and results where a baseline outperforms Transfer are shown in red. Statistical significance ($p < 0.05$) of the transfer model with respect to the baseline is indicated with the symbol †. ‘R-type’ represents resource type (BL:Bilingual-based, ML:Monolingual-based). ‘F-type’ represents feature representation (CW: Cross-lingual Word embeddings, BSW: Bilingual Sentiment Embeddings and Weights, +Lex: With lexicalization, CL: Cross-lingual Word Clusters, SWN: Sentiwordnet.)

line in all cases, and the lexical baseline SWN-rule in all cases but one (Sinhalese(si)) - in fact, for Sinhalese, no other method we tried or compared to beats this baseline. The difference in performance is statistically significant for all languages except Tigrinya, Uyghur, and Sinhalese, which due to relatively small evaluation dataset sizes, require a threshold of roughly ten F-measure points for statistical significance. However, the repeatability of the transfer model's success for these languages under different feature configurations (as shown in Section 4.5.3) is encouraging.

Transfer vs. Supervised.

In general, the transfer model does not lag very far behind the supervised model. The difference between the transfer model and the supervised model ranges from as low as 7 points (Spanish) to as high as 28 points (Slovak) (in fact, Slovak transfers better from other languages, namely its Western Slavic sister language, Polish, as will be shown in Chapter 5), and in one case (Chinese), the transfer model actually surpasses the supervised model. On average, the best sentiment transfer model trails the supervised model by about 15 points, with the the smallest differences observed for Spanish(es), Bulgarian(bg), German(de), Arabic(ar), and Portuguese(pt) (below 10 points) and the highest differences observed for Slovak(sk) and Croatian(hr) (above 20 points). The majority baseline for these last two languages is notably low as well (11.9 and 12.8) due to the skew in their datasets, which contain a small amount of neutral data compared to the English data they were trained on (See Chapter 3 Table 3.3, for sentiment distributions of the datasets). Yet, the transfer model still easily exceeds the baseline for Slovak and Croatian (40.8 and 39.7), indicating that it learns to identify sentiment in the target language.

Resources.

When considering feature and resource representations for the transfer model, it is clear that bilingually trained representations (BL) most often result in the best model, contributing to the best transfer configuration in 14 out of 17 languages,

Best Transfer Model: Conclusion
Transfer model outperforms baseline *(16/17 languages)*
Transfer model lags supervised model by 14.7 points average
Best Embedding Generation: Embeddings built with bilingual context *(14/17)*
Best Features: Lexicalization *(11/17)* and Bilingual Sentiment Weights *(8/17)*
Best Corpora: In-domain (LDC: *9/17*, EP: *6/17*)

while monolingual-based Dict-CS representations result in the best model for Chinese(zh), Sinhalese(si), and Swedish(sv). Furthermore, the best bilingually trained embeddings are most often trained on an in-genre or in-domain parallel corpus (LDC or EP) (15 out of 17 languages). In two cases, Slovak and Croatian, the comparable corpus formed of language-linked Wikipedia articles actually outperforms the parallel corpora alternatives. In fact, the out-of-domain parallel corpus QB does not result in the best model for any language.

Features.

Of the bilingual feature representations considered for training the cross-lingual model, lexicalization (+Lex) and bilingual sentiment embeddings and weights (BSW) contribute most often to the best model: 11 out of 17 languages for lexicalization and 8 out of 17 languages for BSW. On the other hand, the contribution of Sentiwordnet scores (SWN) and cluster embeddings (CL) is less pronounced (4 for CL and 4 for SWN).

We study the results for bilingual feature resources and representations in further detail in the next sections.

4.5.2 Evaluation of Bilingual Feature Resources

Tables 4.4 and 4.5 show the performance of the cross-lingual model when using monolingual-based and bilingual-based word embeddings created using different resources: in-domain and in-genre parallel corpora (Table 4.4), and out-of-domain, comparable, and monolingual corpora (Table 4.5). The first set of results reflects

	Bilingual Resource Evaluation: In-domain and In-genre Parallel									
	LDC					*EP*				
	DICT-CS		VECMAP	MUSE	BL	DICT-CS		VECMAP	MUSE	BL
	-Lex	+Lex	-Lex	-Lex	-Lex	-Lex	+Lex	-Lex	-Lex	-Lex
ar	36.7	30.0	36.6	37.1	**40.4**†	–	–	–	–	–
bg	–	–	–	–	–	24.6	43.5	36.4	39.8	**48.4**$^{\heartsuit\dagger\diamondsuit}$
de	–	–	–	–	–	40.5	45.4	47.6	48.0	**49.2**$^{\diamondsuit}$
es	40.9	42.2	39.0	39.7	42.2$^{\heartsuit\dagger}$	27.3	39.4	38.5	41.7	**43.2**$^{\heartsuit\dagger\diamondsuit}$
fa	34.9	26.5	49.1	48.4	**50.4**$^{\diamondsuit}$	–	–	–	–	–
hu	37.7	42.3	44.6	46.3	47.4$^{\dagger\diamondsuit}$	28.2	31.8	41.2	47.6	**48.2**$^{\dagger\diamondsuit}$
hr	–	–	–	–	–	–	–	–	–	–
pl	–	–	–	–	–	24.6	43.3	28.8	36.8	**43.5**$^{\heartsuit\dagger}$
pt	–	–	–	–	–	26.5	39.3	40.2	**42.1**$^{\heartsuit}$	41.1
ru	43.4	48.1	44.0	47.3	**49.0**†	–	–	–	–	–
si	26.9	35.2	24.7	**36.2**	31.4	–	–	–	–	–
sk	–	–	–	–	–	17.9	20.4	34.6	33.4	**38.4**$^{\heartsuit\dagger\diamondsuit}$
sl	–	–	–	–	–	31.8	40.1	32.5	35.2	**41.4**$^{\heartsuit\dagger}$
sv	–	–	–	–	–	29.4	**49.0**$^{\diamondsuit}$	35.3	34.9	43.6
ti	36.3	36.9	29.0	**37.7**	34.5	–	–	–	–	–
ug	25.4	37.5	30.9	26.7	**45.2**$^{\heartsuit}$	–	–	–	–	–
zh	55.9	56.3	53.0	**58.1**	52.9	–	–	–	–	–
AVG	37.6	39.4	39.0	41.9	43.7	27.9	39.1	37.2	39.9	**44.1**

Table 4.4: Macro-averaged F-measure of cross-lingual model using in-genre (LDC) and in-domain (LDC, EP) parallel corpora with bilingual-based and monolingual-based embedding methods. Best results are shown in **bold**. For each corpus (LDC and EP), statistical significance ($p<0.05$) between BL and corresponding MUSE ($^{\heartsuit}$), VECMAP (†), and the best Dict-CS model ($^{\diamondsuit}$) is indicated. DICT-CS, VECMAP, and MUSE are monolingual-based methods, while BL is a bilingual-based method learned directly on the parallel corpus. '-Lex' means no lexicalization occurs during training, '+Lex' means target language lexicalization occurs during training.

the more desirable scenario where an in-domain or in-genre parallel - albeit possibly small - corpus is available for the target language. The second set of results reflects the scenario where only an out-of-domain parallel corpus, comparable corpus with no dictionary, or monolingual corpus with no dictionary is available.

For *supervised scenarios* (a parallel corpus or bilingual dictionary is available), we show results using the three monolingual-based embedding methods: Dict-CS, vecmap, and muse, along with bilingual-based embedding training bl. Dict-CS is shown using both lexicalization (+Lex) and no lexicalization (-Lex) configurations, while all other cross-lingual models have not been trained using any translation to the target language: i.e, the model relies only on the bilingual nature of the embeddings to represent the target language.

For *unsupervised scenarios* (only a non-parallel corpus and no dictionary is available[12]), we show results using our bilingual-based comparable corpus training bl as well as with vecmap trained on comparable corpora, and finally, only vecmap is used for the scenario where a monolingual corpus and nothing else is available.

4.5.2.1 Discussion

In-genre and In-domain Parallel Corpora.

We first consider in-domain and in-genre parallel corpora. The results in Table 4.4 show that bilingual-based embeddings bl trained directly on EP or LDC, without any lexicalization, generally outperform all the monolingual-based embeddings Dict-CS, vecmap, and muse on identifying sentiment in the target language, even though monolingual-based embeddings are built in the supervised configuration with access to a bilingual dictionary. We can see this as bl outperforms other representations on average and results in the best model in 11 out of 16 languages. The second-best performing model is muse (4 out of 16 languages). In terms of statistical significance, we observe 8 of these languages with bl significantly outperforming vecmap, 6 significantly outperforming muse, 6 significantly outperforming dict-cs *with* lexicalization and almost all languages for dict-cs *without* lexicalization. On the other

[12] Except for the 61 keywords originally used to retrieve the comparable corpora.

hand, monolingual-based methods perform significantly better than BL for only 2 languages: Portuguese(pt) with MUSE, and Swedish(sv) with DICT-CS+LEX.

We note that MUSE uses an external validation dictionary on top of the LDC parallel corpus dictionary.

Effect of Lexicalization on Dict-CS and BL.

We see that DICT-CS representations are greatly improved through translation, with lexicalization leading to an increase in F-measure of 37.6 to 39.4 on average for LDC and 27.9 to 39.1 for EP. On the other hand, the bilingual-based embeddings, particularly because they benefit from the bilingual context of the parallel corpus, are able to stand on their own without any lexicalization. Lexicalization impacts the Europarl corpus more than the in-genre LDC corpus.

Cases where ML-based embeddings outperform BL-based embeddings.

We consider the few cases where monolingual-based representation methods better enable the model to identify sentiment. These include Sinhalese(si) and Chinese(zh), where MUSE results in the best F-measures of 36.2 and 58.1, and Swedish (sv), where DICT-CS+LEX results in the best F-measure of 49.0 as well as the best overall model for Swedish. MUSE and DICT-CS also outperform BL embeddings on Tigrinya, with MUSE resulting in the best F-measure of of 37.7 - although our best overall model for Tigrinya is obtained when we use bilingual sentiment weights, as the next section shows. We can observe from the corpus vocabulary sizes shown in Table 4.2, that Chinese has a large monolingual corpus vocabulary (466.8k) relative to its parallel vocabulary size (6.9k). For most languages with smaller monolingual vocabularies, such as Uyghur, Arabic, Persian, or Bulgarian (See Table 4.2) using the available smaller parallel corpus is clearly a better option. Sinhalese does not follow this pattern. Training embedding models with very large monolingual corpora also becomes cumbersome (as we observed with high-resource languages, namely English and German); if computational memory and resources are not available, embeddings

In-domain Parallel Resources: Conclusion
Best embeddings: Bilingual-based *(11/16 languages)*
Second best embeddings: MUSE (monolingual-based) *(3/16 languages)*
Factors affecting performance: Monolingual vs. Parallel vocabulary size
Reliance on lexicalization: Dict-CS *(yes)*, Bilingual-based *(no)*

cannot be trained quickly and easily.

Out-of-domain, Comparable, and Monolingual Corpora.

With out-of-domain, comparable, and monolingual corpora (Table 4.5), we see that BL-based embeddings still perform best on average, resulting in the best model for 6/17 languages, 3/17 languages with the Quran-Bible corpus and 3/17 languages with the Comparable corpus. However, the results are now more mixed. This points to out-of-domain and non-parallel resources being a varied setting where performance is dependent on the resource and the language.

MUSE is still the second best embedding model, resulting in the best model for 2/17 languages and the second highest average performance, while VECMAP results in the best model for 5/17 languages when using *only* a monolingual corpus. It performs substantially well for some languages, such as Bulgarian(bg), Persian(fa), and Chinese(zh), but substantially poorly for others, such as Sinhalese(si), Slovak(sk), and Slovene(sl).

In terms of statistical significance, our significance testing verifies this varied set of results, with an overall advantage to bilingual-based embeddings. With the Quran-Bible corpus, bilingual-based embeddings are significantly better than VECMAP for 10 languages, than DICT-CS+LEX for 7 languages, and than MUSE for 6 languages. On the other hand, we see significant improvements for MUSE over BL for 4 languages, and for DICT-CS with 3 languages if lexicalization is applied and one language (Slovene(sl)) if it is not. With the comparable corpus, BL significantly outperforms VECMAP for 12 languages. With unsupervised monolingual-based embeddings, where

	Resource Evaluation: Out-of-domain, Comp and Mono							
	QB					*Comparable*		*Monolingual*
	DICT-CS		VECMAP	MUSE	BL	VECMAP	BL	VECMAP
	-Lex	+Lex	-Lex	-Lex	-Lex	-Lex	-Lex	-Lex
ar	26.4	$37.3^{\diamondsuit}$	32.7	$37.9^{\heartsuit}$	34.1	26.6	**$38.9^{\dagger}$**	$37.9^{\clubsuit(Q)}$
bg	24.7	33.0	38.5	40.2	$44.1^{\heartsuit\dagger\diamondsuit}$	32.7	$42.1^{\dagger}$	**$45.7^{\clubsuit(C)}$**
de	36.0	43.5	44.6	45.4	**$47.8^{\heartsuit\dagger\diamondsuit}$**	45.5	44.3	47.0
es	27.0	**$42.6^{\diamondsuit}$**	37.4	41.1	$41.2^{\dagger\clubsuit}$	36.9	36.6	$39.3^{\clubsuit(C)}$
fa	18.4	40.1	$49.1^{\dagger}$	48.1	$48.4^{\diamondsuit}$	47.8	48.9	**50.5**
hu	30.2	41.1	40.1	**$46.0^{\heartsuit}$**	$44.5^{\dagger\diamondsuit}$	37.0	$44.4^{\dagger}$	45.2
hr	18.5	30.8	35.0	38.3	**$38.7^{\dagger\diamondsuit\clubsuit}$**	29.1	$38.4^{\dagger}$	31.9
pl	23.6	**$41.7^{\diamondsuit}$**	29.2	$38.7^{\heartsuit}$	$35.0^{\dagger}$	33.1	$38.4^{\dagger\clubsuit}$	35.6
pt	20.2	38.6	34.8	**$40.6^{\heartsuit}$**	$39.6^{\dagger}$	32.9	$37.3^{\dagger}$	38.9
ru	24.1	$44.8^{\diamondsuit}$	40.7	43.8	42.5	39.4	$44.1^{\dagger}$	**45.6**
si	–	–	–	–	–	24.3	**$31.5^{\dagger\clubsuit}$**	21.5
sk	15.1	22.6	31.5	30.9	$34.1^{\heartsuit\dagger\diamondsuit}$	27.9	**$40.8^{\dagger\clubsuit}$**	23.0
sl	**$35.2^{\diamondsuit}$**	32.2	28.7	31.7	$34.0^{\heartsuit\dagger\clubsuit}$	28.3	$33.3^{\dagger\clubsuit}$	27.4
sv	27.0	**39.1**	33.5	33.0	$37.9^{\heartsuit\dagger\clubsuit}$	24.4	$36.2^{\dagger}$	33.5
ti	–	–	–	–	–	–	–	**29.5**
ug	16.1	30.0	35.7	26.2	**$38.2^{\heartsuit\clubsuit}$**	$33.4^{\dagger}$	28.5	26.3
zh	16.5	30.3	50.5	$55.1^{\heartsuit}$	$44.6^{\diamondsuit}$	25.0	$34.5^{\dagger}$	**$59.6^{\clubsuit(Q)}$**
AVG	23.9	36.5	37.5	39.8	**40.3**	32.8	38.6	37.5

Table 4.5: Macro-averaged F-measure of cross-lingual model using out-of-genre and out-of-domain parallel corpora (QB), comparable corpora, and monolingual corpora with bilingual-based and monolingual-based methods. For each corpus (QB and Comparable), statistical significance (*p<0.05*) between BL and corresponding MUSE ($^{\heartsuit}$), VECMAP (†), and the best DICT-CS model ($^{\diamondsuit}$) is indicated. Statistical significance between Monolingual VECMAP and BL is also indicated: with the symbol $^{\clubsuit}$ if BL >VECMAP, otherwise $^{\clubsuit(Q)}$ for BL-QB and $^{\clubsuit(C)}$ for BL-comparable. DICT-CS, VECMAP, and MUSE are monolingual-based methods, while BL is a bilingual-based method learned directly on the bilingual corpus. '-Lex' means no lexicalization occurs during training, '+Lex' means target language lexicalization occurs during training. Comparable and monolingual corpora are 'unsupervised'; no dictionary or parallel corpus is available.

VECMAP has the strongest advantage, we observe 4 languages (Arabic(ar), Bulgarian(bg), Spanish(es), and Chinese(zh)), where monolingual VECMAP embeddings outperform a bilingual-based model, either using out-of-domain (BL-QB) or comparable (BL-comparable) corpora. On the other hand, we observe significant improvements of bilingual-based embeddings over unsupervised VECMAP for 9 languages in total: 5 languages (Spanish(es), Croatian(hr), Slovene(sl), Swedish(sv), and Uyghur(ug)) if using the out-of-domain corpus and 4 languages (Polish(pl), Sinhalese(si), Slovak(sk), and Slovene(sl)) if using the comparable corpus.

Monolingual Corpus Vocabulary.

Languages with large monolingual corpus vocabulary sizes compared to their Quran-Bible corpus (Table 4.5: 466.8K to 27.3K for Chinese(zh), 388.5K to 18.6K for Hungarian(hu), 543.7K to 50.3K for Russian) tend to do well with VECMAP or MUSE. This effect is more pronounced with the out-of-domain corpus compared to the in-domain LDC and EP corpora, where smaller corpora were more likely to be sufficient to produce better sentiment results. Arabic(ar), Slovene(sl), and Slovak(sk), on the other hand, which also have small QB vocabularies (17.3K, 14.2K and 16.2K) but monolingual corpora of 150K and 190K, perform better with BL. Finally, for languages like German(de), which is richly resourced for all corpora, BL performs the best, but the differences in results among different methods are not substantial.

Effect of Lexicalization.

DICT-CS with lexicalization results in the best model for Spanish(es), Polilsh(pl), and Swedish(sv), 3/17 languages. However, without lexicalization, its average performance drops from 36.5 to 23.9. This effect is most pronounced for the Quran-Bible corpus compared to all available parallel corpora.

Performance of Comparable vs. Quran-Bible Corpus.

If we compare the use of the bilingually trained comparable corpora embeddings to the Quran and Bible parallel corpus embeddings, we see that comparable corpus training with BL comes quite close, and sometimes even outperforms parallel corpus training using the Quran and Bible corpus. This is the case for 5 languages: Arabic (38.9 vs. 34.1 F-measure), Persian (48.9 vs. 48.4), Polish (38.4 vs. 34.0), Russian (44.1 vs. 42.5), and Slovak (40.8 vs. 34.0). The greatest improvements are observed for Arabic(ar), Polish(pl), and Slovak(sk), all of which don't have substantially large out-of-domain and monolingual vocabularies, and for Sinhalese(si), which has no QB corpus; thus, an in-domain comparable corpus using bilingual-based embeddings is preferable for sentiment transfer when a *small* in-domain parallel corpus is not available or when a *large* out-of-domain parallel corpus or large monolingual corpus is not available.

Moreover, training bilingual-based embeddings on the comparable corpus outperforms both DICT-CS and VECMAP embeddings created with QB supervision, as well as unsupervised VECMAP embeddings created using the same comparable corpus. These results are notable, indicating that a comparable corpus is effective for training cross-lingual sentiment models without any translation dictionary, and demonstrates once again that the content of the corpus and its similarity in domain to the evaluation data is an important factor for cross-lingual sentiment analysis. Our results are also consistent with the comparability and sentiment comparability scores computed on the comparable and QB corpora in Chapter 3, which showed comparable corpora having equal or higher comparability to the QB corpus.

Performance of Unsupervised Monolingual Corpus.
This scenario refers to the last column in Table 4.5. It is interestingly, the best out of all possible options for Chinese(zh). For Bulgarian(bg), Russian(ru), and Persian(fa), using VECMAP embeddings under completely unsupervised settings outperforms both out-of-domain parallel corpora and comparable corpora options - although using an in-domain parallel corpus still works better for these three languages, as shown in Tables 4.3 and 4.4. For Arabic(ar), Sinhalese(si), and Slovak(sk), using the com-

Out-of-domain, Comparable, and Monolingual Resources: Conclusion

Varied language-dependent resources

Factors: In-domain vs. Out-of-domain, Vocabulary size

Best average model: Bilingual-based embeddings *(6/16 languages)*

Second best average model: MUSE (monolingual-based) *(2/16 languages)*

Monolingual VECMAP: best model for *5/17 languages*

Comparable *(5 outperform)* vs. Out-of-domain *(9 outperform)* for bilingual-based

Reliance on lexicalization: Dict-CS *(yes)*, bilingual-based *(no)*

Languages benefiting from VecMap and MUSE:

Bulgarian, Persian, Hungarian, Portuguese, Russian, Chinese

Languages benefiting from bilingual-based embeddings:

Arabic, German, Croatian, Sinhalese, Slovak, Uyghur

Languages benefiting from code-switched embeddings:

Spanish, Polish, Swedish

parable corpus with BL is a preferable option, and the remaining languages with Quran-Bible available do better with the out-of-domain corpus. As for Tigrinya(ti), where a comparable corpus is not available, monolingual training is the best option if no parallel corpus is available at all, resulting in an F-measure of 29.5. However, with a vocabulary of 5.4K (*only 11.8K translation sentences*) and bilingual sentiment embeddings, the small parallel corpus yields notable improvements with an F-measure of 39.5 without lexicalization (see next section) and 40.9 with lexicalization. This is a notable result, as a small parallel corpus is usually as likely to be available for a low-resource language as a very large, digitally available monolingual corpus.

4.5.3 Evaluation of Bilingual Feature Representations

We turn to the evaluation of bilingual feature representations. Tables 4.6 and 4.7 show the cross-lingual model's performance using bilingually trained word embeddings on their own (CW), bilingually trained word embeddings with Sentiwordnet scores (SWN), and bilingually trained sentiment embeddings and weights (BSW) on all translation corpora (LDC, EP, QB, and Comparable). Lexicalization occurs in supervised scenarios, and as SWN relies on translating the Sentiwordnet lexicon to

	Feature Representation Evaluation 1									
	LDC					*EP*				
	CW		SWN	BSW		CW		SWN	BSW	
	-Lex	+Lex	+Lex	-Lex	+Lex	-Lex	+Lex	+Lex	-Lex	+Lex
ar	40.4	45.2	43.2	**45.9**†	44.7	–	–	–	–	–
bg	–	–	–	–	–	48.4	49.1	**49.3**	48.5	47.9
de	–	–	–	–	–	**49.2**	**49.2**	48.7	48.8	49.1
es	42.2	**44.4**†	44.3	42.0	44.3	43.2	42.4	42.7	42.3	42.5
fa	50.4	52.2	51.6	50.0	**52.9**$^{\heartsuit}$	–	–	–	–	–
hu	47.4	48.7	48.6	45.6	**49.1**$^{\heartsuit}$	48.2	47.0	46.0	48.5	46.8
hr	–	–	–	–	–	–	–	–	–	–
pl	–	–	–	–	–	43.5	40.8	40.9	**43.9**	40.6
pt	–	–	–	–	–	41.1	42.1	41.7	41.5	**42.5**†
ru	49.0	49.9	49.2	49.3	**50.2**	–	–	–	–	–
si	31.4	34.2	34.0	32.5	**34.7**	–	–	–	–	–
sk	–	–	–	–	–	38.4	**40.2**†	38.8	39.2	38.8
sl	–	–	–	–	–	41.4	42.0	41.7	40.0	**42.3**$^{\heartsuit}$
sv	–	–	–	–	–	43.6	45.8	46.1	42.9	**46.3**†
ti	34.5	36.4	36.6	39.7	**40.9**	–	–	–	–	–
ug	**45.2**	40.9	39.4	43.4	41.4	–	–	–	–	–
zh	52.9	52.0	51.8	**53.5**	48.8	–	–	–	–	–
AVG	43.7	44.9	44.3	44.7	**45.2**	44.1	44.3	44.0	44.0	44.1

Table 4.6: Macro-averaged F-measure of cross-lingual model using in-genre (LDC) and in-domain (LDC, EP) parallel corpora with bilingual feature representations.‘CW’ are cross-lingual word embeddings learned on a bilingual corpus, ‘SWN’ adds Sentiwordnet scores to CW, ‘BSW’ are bilingual sentiment embeddings and weights. ‘-Lex’ means no lexicalization occurs during training, ‘+Lex’ means target language lexicalization and BSW weight update occurs during training. Statistical significance (p <*0.05*) of the model with the best representation feature(s) with respect to the corresponding model with no added feature (CW) is indicated with the symbol †. Mild significance (p <*0.08*) is indicated with the symbol $^{\heartsuit}$. All experiments are run 5 times and the averaged result is presented.

the target language, it is only available during this configuration.

Bilingual Feature Representations for In-domain Corpora.
We observe that for in-domain and in-genre corpora (Table 4.6), 11 out of the 17 cross-lingual models perform best when using bilingually trained sentiment embeddings and weights, and similarly, 11 out of the 17 cross-lingual models result in the best performance after applying target language lexicalization, while only one model (Bulgarian(bg)) obtains the best result using SWN scores. BSW outperforms SWN consistently, which indicates that learning sentiment context bilingually is more helpful than learning sentiment scores in the source language and then projecting them to the target language.

Together, our approach for pre-training bilingual sentiment embeddings and updating the weights during training by allowing the training data to be lexicalized, results in the best overall performing model. This method (BSW+Lex) results in the best method for Persian(fa), Portuguese(pt), Russian(ru), Sinhalese(si), Slovene(sl), Swedish(sv), and Tigrinya(ti). On the other hand, BSW alone results in the best method for Arabic(ar), Polish(pl), and Chinese(zh), while lexicalization alone results in the best method for Spanish(es) and Slovak(sk). Arabic for example benefits from both bilingually trained sentiment embeddings and lexicalization, but their combination results in no significant improvement.

In terms of statistical significance, we observe that models having BSW+Lex as the best feature are significant for 5 languages (2 strongly significant: Portuguese(pt) and Swedish(sv)), and 3 mildly significant (Persian(fa), Hungarian(hu), and Slovene(sl)). Models using lexicalization as the best feature are significant for 2 languages (Spanish(es) and Slovak(sk)), and models using BSW as the best feature are significant for 1 language (Arabic(ar)). On the other hand, models using SWN+Lex as the best feature are not significant for any language and models where additional features do not help are not significant (German(de) and Uyghur(ug)). Considering dataset sizes for Tigrinya, Uyghur, and Sinhalese, we have found that a significance threshold of roughly 10 F-measure points is needed for statistical significance to be observed with these languages.

Effect of Corpus on Pre-trained Sentiment Embeddings.
In general, pre-trained bilingual sentiment embeddings are most helpful when using the LDC corpus; with Europarl, the impact is not as consistent - particularly when BSW is not updated during training. This could be because the sentiment content of the Europarl corpus is not as well-suited for pre-training sentiment scores as the LDC corpus; as shown in Chapter 3 (Table 3.12), the proportion of subjective-neutral words (e.g 'alliance', 'assessment') tagged by the MPQA lexicon is quite high with Europarl (33.6%) in comparison to the LDC corpora (23.8%), Wikipedia corpora (28.9%), and religious corpora (19.5%). On the other hand, both the LDC and Wikipedia corpora are more evenly distributed amongst the three sentiment labels.

Only two languages don't benefit from any additional feature representation beyond bilingual embeddings under in-domain corpora: German(de) and Uyghur(ug). Both languages have a high proportion of neutral labels and a lower proportion of negative sentiment labels in their datasets; we investigate sentiment distribtuion of test datasets in the error analysis section.

Bilingual Feature Representations for Out-of-domain and Comparable Corpora.
With out-of-domain parallel corpora and comparable corpora (Table 4.7), the results are mixed across Quran-Bible and comparable corpora, but bilingual sentiment weights with lexicalization (BSW+Lex) is still the best performing feature on average. Overall, BSW results in the best model for 9 out of 16 languages in this setting, lexicalization results in the best model for 8 out of 16 languages, while BSW+Lex results in the best model for 3 out of 16 languages. (The reason this number is now

Features for In-domain Corpora: Conclusion Factors: Pre-training Corpus, Sentiment in Test Data
Best feature: BSW+Lex *(7/17 languages)* *(Bilingual sentiment embeddings with lexicalization)* Second best feature: BSW*(3/17 languages)* *(Bilingual sentiment embeddings)* Third best feature: CW+Lex *(2/17 languages)* *(Lexicalization)*
Languages benefiting from BSW+lex: *Persian, Portuguese, Russian, Sinhalese, Slovene, Swedish, Tigrinya* **Languages benefiting from BSW:** *Arabic, Polish, Chinese* **Languages benefiting from CW+Lex:** *Spanish, Slovak* **Languages not benefiting from any feature:** *German, Uyghur*

lower is because 5 languages did better with the comparable corpus, which does not have a lexicalization configuration). As to SWN+Lex, it results in the best model for 2 out of 16 languages, Spanish(es) and Uyghur(ug), but it is again outperformed by BSW+Lex in the majority of cases.

In terms of statistical significance, we observe that with the Quran-Bible corpus, BSW+Lex results in significant improvements for 4 languages where it is the best feature (Arabic(ar), Spanish(es), Polish(pl), and Slovene(sl)), lexicalization results in significant improvements for 2 languages where it is the best feature (Portuguese(pt) and Swedish(sv)), while BSW results in mildly significant improvement for one language where it is the best feature (Bulgarian(bg)). On the other hand, SWN+Lex results in significant improvement for one language where it is the best feature (Spanish(es)). Languages where no feature improvement occurs (German(de)) show no significance. These results are consistent with what was observed for in-domain corpora. For comparable corpora, we observe 3 languages where BSW results in significant or mildly significant improvement (Arabic(ar), Swedish(sv), and Chinese(zh)); with the remaining 4 languages where BSW results in improvement, we do not observe statistically significant differences. For 4 other languages, not

	Feature Representation Evaluation 2						
	QB					*Comparable*	
	CW	CW	SWN	BSW	BSW	CW	BSW
	-Lex	+Lex	+Lex	-Lex	+Lex	-Lex	-Lex
ar	34.1	37.7	38.3	35.1	38.6†	38.6	**39.2**$^{\heartsuit}$
bg	44.1	44.1	43.3	**45.7**$^{\heartsuit}$	44.0	42.1	42.4
de	**47.8**	46.5	46.7	47.0	47.7	44.3	44.4
es	41.2	42.7	**43.0**†	40.2	**43.0**†	36.6$^{\heartsuit}$	35.6
fa	48.4	**50.9**	50.1	49.2	50.7	48.9†	45.2
hu	44.5	40.9	40.0	43.2	40.6	44.4	**45.1**
hr	38.7	36.9	37.4	37.4	36.1	38.4	**39.7**
pl	35.0	36.5	36.5	35.7	**38.5**†	38.4†	37.3
pt	39.6	**41.3**†	41.2	39.5	41.2	37.3	36.7
ru	42.5	39.5	38.8	41.1	39.4	44.1	**44.3**
si	–	–	–	–	–	31.5	**32.0**
sk	34.1	34.2	35.2	32.8	33.6	**40.8**†	34.9
sl	34.0	37.2	36.8	35.4	**37.7**†	33.3	34.1
sv	37.9	**42.1**†	41.3	41.1	41.7	36.2	38.2$^{\heartsuit}$
ti	–	–	–	–	–	–	–
ug	38.2	37.5	**40.2**	38.5	39.9	28.5	26.0
zh	44.6	**47.8**	44.3	42.0	47.2	34.5	41.5†
AVG	40.3	41.1	40.9	40.3	**41.3**	38.6	38.5

Table 4.7: Macro-averaged F-measure of cross-lingual model using out-of-genre and out-of-domain parallel corpora (QB) and comparable corpora with bilingual feature representations. ‘CW’ are cross-lingual word embeddings learned on a bilingual corpus, ‘SWN’ adds Sentiwordnet scores to CW, ‘BSW’ are bilingual sentiment embeddings and weights. ‘-Lex’ means no lexicalization occurs during training, ‘+Lex’ means target language lexicalization and BSW weight update occurs during training. For each of QB and comparable corpora, statistical significance ($p < 0.05$) of the model with the best representation feature(s) with respect to the corresponding model with no added feature (CW) is indicated with the symbol †. Mild significance ($p < 0.08$) is indicated with the symbol $^{\heartsuit}$. Comparable corpora are ‘unsupervised’; no dictionary or parallel corpus is available. All experiments are run 5 times and the averaged result is presented.

Features for Out-of-domain and Comparable Corpora: Conclusion
Varied, language-dependent resources
Factors: Pre-training Corpus, Sentiment in Test Data

Best average feature: BSW+Lex
(Bilingual sentiment embeddings with lexicalization)
Second best average feature: CW+Lex
(Lexicalization)
Comparable *(6 best)* vs. Out-of-domain *(10 best)*

Languages benefiting from BSW+lex:
Spanish, Polish, Slovene
Languages benefiting from BSW:
Arabic, Bulgarian, Hungarian, Croatian, Polish, Russian, Sinhalese
Languages benefiting from CW+Lex:
Persian, Portuguese, Swedish, Chinese
Languages not benefiting from any feature:
German

using BSW is significantly or mildly significantly better. BSW therefore has varied results among languages when using comparable corpora.

Performance of Comparable vs. Quran-Bible Corpus.

QB results in the best model for 11 out of 17 languages, while the unsupervised Comparable corpus results in the best model for 6 out of 17 languages: Arabic(ar), Hungarian(hu), Croatian(hr), Russian(ru), Sinhalese(si), and Slovak(sk). Among these, 5 of the languages use BSW. Results with comparable corpora are once again encouraging considering the unsupervised nature of this setting.

Performance of Cross-lingual Cluster Features.

Because our initial experiments with cross-lingual clusters, although not detrimental, did not yield notable improvements generally, we did not pursue them in our final bilingual-based models. Moreover, they did not result in the best overall model for any language (as shown in Table 4.3 at the onset of this section) except for Persian(fa) with the LDC corpus. Table 4.8 shows results using cluster features (CL) with LDC, EP, and QB corpora. We observe 9 languages overall that result in an improvement with clusters while 8 languages don't. While small improvements are obtained using cross-lingual clusters when using the LDC corpus - with greater improvements observed for Persian(fa) and Tigrinya(ti), the improvement overall and for most languages is not as prominent as what was observed with bilingual sentiment embeddings or with lexicalization. In fact, our targeted sentiment experiments in Chapter 6 also show that the effect of word clusters on sentiment identification is not so definitive (they are more helpful for target entity identification). These results probably have to do with the fact that semantic word vector clusters do not always capture synonymic relations; the same cluster may include words having antonymic relationships, such as 'wonderful' and 'awful', which would affect the performance of detecting sentiment.

4.5.3.1 Evaluating Topically-Aligned Comparable Corpora

This part examines cross-lingual model results using bilingual embeddings created from topically-aligned comparable corpora, as opposed to our main article-aligned comparable corpus. Table 4.9 shows the performance of the topically-aligned corpus using 'length-ratio-shuffle' with monolingual ordering (Column 2), 'length-ratio-shuffle' with merge-TFIDF (Column 3), and the article-aligned corpus (Column 4). The majority baseline is shown in Column 1. Clearly, using an article-aligned corpus is more beneficial across target languages. However, the topically aligned corpus outperforms the majority baseline and in addition, Merge-TFIDF outperforms Merge with monolingual ordering for most languages.

4.5.4 Comparison with Previous Work

We compare our cross-lingual model with the adversarial transfer model of Chen et al. (2018), which is publicly available[13]. We ran the adverserial model on the English training data for 5 epochs with our standard bilingual embeddings and used their default configurations for the remaining hyperparameters. Since this model requires

[13] https:// github.com/ccsasuke/adan

	Cluster Feature Evaluation					
	LDC		*EP*		*QB*	
	CW	CL	CW	CL	CW	CL
ar	**40.4**	40.0	–	–	34.1	33.6
bg	–	–	**48.4**	47.6	44.1	44.3
de	–	–	**49.2**	48.5	47.8	47.5
es	42.2	42.2	42.4	**43.6**	41.2	41.5
fa	50.4	**53.0**	–	–	48.4	48.6
hu	47.4	48.1	48.2	**48.9**	44.5	44.2
hr	–	–	–	–	38.7	**38.1**
pl	–	–	**43.6**	42.1	35.0	34.9
pt	–	–	**41.1**	40.5	39.6	39.4
ru	49.0	**49.7**	–	–	42.5	44.0
si	31.4	**31.8**	–	–	–	–
sk	–	–	38.4	**39.3**	34.1	33.4
sl	–	–	**41.4**	40.7	34.0	34.4
sv	–	–	43.6	**45.1**	38.0	39.6
ti	34.5	**36.5**	–	–	–	–
ug	**45.2**	43.5	–	–	38.2	36.7
zh	**52.9**	52.6	–	–	44.6	49.5
AVG	43.7	**44.2**	44.1	44.0	40.3	40.6

Table 4.8: Macro-averaged F-measure of cross-lingual model using in-genre (LDC) and in-domain (LDC, EP) and out-of-domain (QB) parallel corpora with bilingual cluster features. 'CW' are cross-lingual word embeddings learned on a bilingual corpus, while 'CL' adds cross-lingual cluster embeddings to CW. All experiments are run 5 times and the averaged result is presented.

unlabeled target data, we used the validation dataset splits created from the dataset of Mozetič et al. (2016) for the languages that had them and we used unlabeled LDC monolingual data for Uyghur, Tigrinya, Chinese, and Sinhalese. Table 4.10 shows the results.

Our direct cross-lingual model when using the same bilingual embeddings outper-

	Comparable Corpus Evaluation			
	Majority	*Topic-Aligned*		*Article-Aligned*
		Merge	Merge-TFIDF	
ar	18.6	20.6	21.1	**38.6**
bg	22.4	26.8	28.6	**42.1**
de	23.9	32.4	37.7	**44.3**
es	19.3	26.2	27.7	**36.6**
fa	17.9	18.2	20.7	**48.9**
hu	16.5	23.4	22.4	**44.4**
hr	12.8	14.6	13.9	**38.4**
pl	13.8	18.4	30.2	**38.4**
pt	17.3	21.4	22.2	**37.3**
ru	20.0	27.6	25.3	**44.1**
si	16.2	16.2	21.5	**31.5**
sk	11.9	28.6	24.6	**40.8**
sl	20.6	25.3	25.3	**33.3**
sv	16.1	26.9	31.1	**36.2**
ti	–	–	–	—
ug	23.6	25.1	25.7	**28.5**
zh	19.7	24.3	32.7	**34.5**
AVG	18.2	23.5	25.7	**38.6**

Table 4.9: Macro-averaged F-measure of direct transfer cross-lingual model with neutral majority baseline and comparable corpus embeddings built with topically-aligned corpora ('Merge' and 'Merge-TFIDF') and article-aligned corpora. Best results are shown in **bold** and the best topic-aligned result is hown in blue.

	Adversarial Model Comparison							
	LDC		*EP*		*QB*		*Comp*	
	Adv	DT	Adv	DT	Adv	DT	Adv	DT
ar	34.6	**40.4**	–	–	37.3	34.1	31.9	38.6
bg	–	–	36.2	**48.4**	34.5	44.1	35.9	42.1
de	–	–	44.9	**49.2**	44.4	47.8	42.1	44.3
es	40.3	42.2	39.6	**43.2**	39.4	41.2	36.2	36.6
fa	32.6	50.4	–	–	30.3	48.4	29.2	**48.9**
hu	43.4	47.4	41.5	**48.2**	43.0	44.5	42.1	44.4
hr	–	–	–	–	37.3	**38.7**	34.9	38.4
pl	–	–	40.2	**43.5**	37.6	35.0	41.6	38.4
pt	–	–	37.8	**41.1**	37.6	39.6	33.8	37.3
ru	37.8	**49.0**	–	–	40.2	42.5	42.9	44.1
si	30.7	**31.4**	–	–	–	–	28.2	31.5
sk	–	–	40.3	38.4	35.4	34.1	35.4	**40.8**
sl	–	–	41.1	**41.4**	39.6	34.0	35.8	33.3
sv	–	–	41.1	**43.6**	40.6	37.9	39.0	36.2
ti	27.4	**34.5**	–	–	–	–	–	–
ug	30.2	**45.2**	–	–	32.7	38.2	32.8	28.5
zh	34.3	**52.9**	–	–	30.2	44.6	35.5	34.5
AVG	34.6	43.7	40.3	**44.1**	37.3	40.3	36.1	38.6

Table 4.10: Macro-averaged F-measure of direct transfer cross-lingual model (DT) with bilingual-based embeddings, and adversarial transfer model (Adv) using direct in-genre (LDC) and in-domain (LDC, EP) corpora, out-of-domain parallel corpora (QB), and comparable corpora (Comp). All experiments are run 5 times and the averaged result is presented. Best results for the language are shown in bold and results where Adv outperforms DT are shown in red.

forms the adversarial model in most configurations, with some exceptions when using QB and comparable corpora, which are highlighted in red in Table 4.10. Comparing across all corpora, DT results in the best model for all 17 languages. Differences in performance between the two models are most apparent when using in-domain and in-genre corpora (LDC), followed by in-domain corpora (EP), and are smaller when using out-of-domain (QB) and comparable corpora, where the adversarial model outperforms DT for 5 languages (including Polish(pl), Slovene(sl), and Swedish(sv)) with each of the two corpora. It is not clear why these languages do better with the adversarial model; as with previous results, the target language vocabulary size of the Quran-Bible or comparable corpus may have played a role. However, the direct transfer model still outperforms overall and on average.

The results demonstrate that a direct transfer model with effective pre-trained embeddings can outperform an adversarially trained model that uses the same embeddings. We note that the adversarial model uses a convolutional neutral network (CNN) while ours uses a bidirectional Long Short-Term-Memory Network (biLSTM). However, Chen et al. (2018) report CNN and biLSTM with attention as their top models, both of which outperform the standard version of biLSTM. It is also possible that the adversarial model requires a larger number of training epochs to achieve better results; however, we have used 5 epochs, the same used for training our model.

4.6 Error Analysis

In order to understand why certain bilingual features helped improve cross-lingual performance in some target languages but not in others, we studied the output of our cross-lingual model on the following languages: Arabic(ar), which benefits from bilingual sentiment embeddings, Spanish(es), which benefits from lexicalization, Slovak(sk), which benefits from comparable corpus training, and German(de), which does not benefit from additional representation features. These languages also represent different language families: Afro-Asiatic(ar), Slavic(sk), Romance(es), and Germanic(de).

Table 4.11 shows the breakdown of sentiment performance by each class for the above languages as well as Tigrinya(ti), which benefits from bilingual sentiment embeddings and lexicalization. The model with the added feature or corpus is shown compared with the alternative model, with scores reflecting Accuracy (acc), Macro-averaged F-Measure (F-Macro), positive sentiment F-Measure (F-Pos), negative sentiment F-Measure (F-Neg), and neutral sentiment F-Measure (F-Neut).

	+Feature/Corpus					-Feature/Corpus				
	Acc	*F-Macro*	*F-Pos*	*F-Neg*	*F-Neut*	*Acc*	*F-Macro*	*F-Pos*	*F-Neg*	*F-Neut*
ar	**48.2**	**46.0**	**37.8**	**44.4**	55.8	46.0	40.7	29.9	36.0	**56.1**
ti	**49.8**	**40.9**	**18.3**	**49.8**	**54.7**	42.7	34.5	12.4	37.1	54.0
es	49.4	**44.4**	**46.4**	**28.2**	58.7	**49.7**	42.3	40.9	25.0	**60.9**
de	57.2	48.8	46.3	31.2	68.8	**57.8**	**49.2**	**46.8**	**31.6**	**69.2**
sk	**40.8**	**40.8**	**41.4**	**38.5**	**42.2**	38.8	38.4	36.1	38.1	41.0

Table 4.11: Accuracy, F-Measure, and breakdown of F-Measure for positive (F-Pos), negative (F-Neg), and neutral (F-Neut) classes with and without added feature/corpus. Feature/Corpus added are respectively BSW vs. CW (ar), BSW+Lex vs. CW (ti), CW+Lex vs. CW (es), BSW vs. CW (de), and Comparable vs. EP (sk). Results are averaged over multiple runs.

We can see that for languages benefiting from additional bilingual feature representations (BSW for Arabic(ar), BSW+Lex for Tigrinya(ti), and CW+Lex for Spanish(es)) (*Rows 1-3*), the added feature (BSW, BSW+Lex, or +Lex) results in a substantial increase in performance on predicting positive and negative sentiment labels in the target language (increase in *F-Pos* from 29.9 to 37.8 for Arabic and *F-Neg* from 37.1 to 49.8 for Tigrinya, for example), while performance on the neutral class is less affected. On the other hand, for German(de), which does not benefit from the addition of BSW or any feature, adding BSW results in a slight drop in performance for all sentiment classes (*Row 4*). BSW does not help German because it leads to a slight drop in precision as the model becomes more aggressive in predicting positive and negative sentiment, leading to more false positives. With Slovak (*Row 5*), we see that using the comparable corpus rather than the Europarl corpus leads to an increase in F-Measure across the board, but in especially for the positive and negative classes (*F-pos* and *F-neg*).

Table 4.12 shows the distribution of the test set among sentiment labels for each of these languages, as well as the English training data. We can see that the distribution of most of the target languages across sentiment labels diverges significantly from that of the training data, with German(de) perhaps the closest in distribution to English as well as the lowest in occurrence of positive and negative sentiment. This could explain why it doesn't benefit from the additional features that increase recall for positive and negative labels. Arabic(ar) and Tigrinya(ti), which benefit most from sentiment embeddings, have the highest proportion of negative sentiment, which is what we would expect in the scenario where a disaster incident occurs in the target language-speaking region: i.e, a considerable proportion of negative sentiment in the evaluation data.

	Test Set Distribution		
	%Pos	*% Neg*	*% Neut*
ar	24.8	36.4	38.8
ti	10.9	36.0	53.1
es	47.6	11.4	40.8
de	25.7	18.4	56.0
sk	52.7	25.6	21.7
en	28.9	24.9	46.3

Table 4.12: Distribution amongst sentiment labels in target test datasets and English training dataset (Pos:positive, Neg:negative, Neut:neutral).

Table 4.13 shows examples of the output of the best model and the alternative model on predicting sentiment in the four target languages: Arabic, Spanish, German, and Slovak. (Tigrinya is not shown because of the lack of access to a native speaker or an available machine translation system for the evaluation output at the time of writing.) Bilingual sentiment scores $v_{sentiment}$ for the models that use BSW (best model for Arabic, alternative model for German) are shown in Table 4.14.

We can see from the Arabic examples that the pre-trained bilingual sentiment

	Input Sentence and Translation	*+Feature/Corpus*	*-Feature/Corpus*	Gold
ar		**BSW**	**CW**	
	مجرد اثارة جدل merely inciting argument	negative	neutral	negative
	اليونسف : هدنة اليمن تمنح الامل مجددا Unicef: Yemen ceasefire gives hope again	positive	neutral	positive
es		**CW+Lex**	**CW**	
	maduramos con los daños, no con los años. we mature with the damage, not with the years	neutral	negative	neutral
	lo conseguiré verás ! ! ! ! jajajjaa I'll get it you'll see! ! ! ! hahahhaa	positive	negative	positive
de		**BSW**	**CW**	
	ich will ins bett ! I want to go to bed!	positive	neutral	neutral
	lass es mich werden let me become it	positive	negative	neutral
	nachts wird echt dunkel hier at night it gets really dark here	negative	neutral	negative
sk		**CW (Comp)**	**CW (EP)**	
	slovensko má streleckého majstra sveta ! slovakia has a shooting world champion!	positive	neutral	positive
	v afrike vznikajú stále nové ohniská eboly new outbreaks of Ebola are emerging in Africa	negative	neutral	negative

Table 4.13: Example outputs with and without added feature/corpus for Arabic, Spanish, German, and Slovak. Feature/Corpus added are respectively BSW vs. CW (ar), CW+Lex vs. CW (es), BSW vs. CW (de), and Comparable vs. EP (sk).

weights indeed enabled the model to better recognize positive and negative Arabic tweets. For example, the input tweet مجرد اثارة جدل *'merely inciting argument'* is correctly classified as negative by the cross-lingual model which uses BSW, but as neutral by the model that only uses CW. The word اثارة *('AvArt')*, which means 'creating', 'inciting', or 'mobilizing' has a negative connotation in Arabic. Relying only on translating the word to English would not have been sufficient for the model to detect this. However, as shown in Table 4.14, the bilingual sentiment weights are able to detect the negative polarity from context. Similarly, the model which uses sentiment embeddings is able to assign positive sentiment to the sentence 'Yemen

	Input Sentence	Bilingual Sentiment Scores		
		positive	*negative*	*neutral*
ar	مجرد merely	0.07057	-0.005	0.07736
	اثارة inciting	-0.07862	**0.15615**	0.03528
	جدل argument	-0.17612	**0.32258**	0.07099
ar	اليونسف Unicef	–	–	–
	: :	0.01212	0.01588	-0.02370
	هدنة ceasefire	**0.14275**	-0.04304	0.01867
	اليمن Yemen	0.06436	0.09848	-0.07511
	تمنح gives	**0.20980**	-0.11760	0.00636
	الامل hope	**0.28353**	-0.09504	-0.13545
	مجددا again	0.06756	0.08647	-0.01710
de	**ich** I	0.05575	-0.06209	0.03337
	will want	0.05316	-0.05173	0.01375
	ins into the	0.00067	0.01995	0.05527
	bett bed	–	–	–
	! !	0.05397	0.00597	0.00828
de	**nachts** nights	0.02549	**0.13300**	0.05235
	wird becomes	**0.11688**	-0.01456	-0.04530
	echt really	**0.24405**	-0.07326	-0.01338
	dunkel dark	–	–	–
	hier here	0.01467	0.00640	0.06755
	.. ..	0.1011	**0.1274**	-0.00965

Table 4.14: Bilingual sentiment scores $v_{sentiment}$ for the examples which use BSW in Table 4.13. Scores for out-of-vocabulary words are represented by dashes '–'.

ceasefire gives hope again', while the alternative model incorrectly classifies it as neutral.

For Spanish, we see that the lexicalized model (CW+Lex) is able to identify the difficult first example in row 2 (Table 4.13) as neutral, while the basic model mistakenly classifies it as negative, likely mislead by the word 'damage'. (The model which uses BSW makes the same error.)

For German, the best model makes a better prediction on the first example ('I want to go to bed'), correctly classifying it as neutral. However, both models make an error on the second example 'let me become it', whose gold label is neutral. The third example is correctly classified as negative by BSW, while the best model mistakenly classifies it as neutral. This supports our conclusion that BSW is better at recalling positive and negative labels in the target language, while it may over-predict when the evaluation data contains fewer instances of sentiment. The bilingual sentiment weights for the German examples are shown in rows 3 and 4 of Table 4.14. We can see that $v_{sentiment}$ is generally of smaller magnitude compared with that of the Arabic examples, likely because of the larger amount of neutral content in the EuroParl corpus which contributes to the sentiment weights when learning embeddings.

For Slovak, we observed, as is reflected in the examples, that the model pre-trained on the comparable corpus contained many more positive and negative predictions than the model pre-trained on the Europarl corpus. This is another instance where the neutral content of the Europarl corpus may have influenced the output, leading to a larger number of out-of-vocabulary words that have positive or negative sentiment.

4.7 Conclusion

This chapter presented both novel methods and extensive experimental analyses for transferring sentiment cross-lingually from English to a target language. The methods that we presented included an approach for pre-training sentiment embeddings and weights bilingually on an appropriate translation corpus, using only a source-language sentiment lexicon. Additionally, the weights may be updated during training by lexicalizing or partially translating the training data into the target language. We also presented an effective strategy for leveraging non-parallel comparable corpora for pre-training bilingual embeddings and sentiment embeddings under unsupervised conditions, which allows the cross-lingual model to be trained using non-parallel bilingual representation features.

The experimental analyses that we presented included a comparison of the performance of different bilingual-based and monolingual-based cross-lingual embeddings created using different resources under supervised and unsupervised conditions: in-domain and in-genre parallel corpora, out-of-domain parallel corpora, contemporary comparable corpora, and purely monolingual corpora, as well as an extensive feature analysis of the contribution of different bilingual representation features: bilingual sentiment embeddings, lexicalization, and bilingual sentiment embeddings with weight update through lexicalization.

Our results allow us to draw several conclusions about the varied conditions tested for in our cross-lingual sentiment analysis experiments:

- **Best Transfer Model.** The cross-lingual transfer model, in its best configuration for each target language, outperforms all baselines for 16 out of 17 languages and comes within acceptable range of a supervised model trained on the same language. The embedding generation method resulting in the best configuration (14/17 languages) was bilingual-based embeddings, with lexicalization and bilingual sentiment embeddings resulting in the best representation features. The most effective corpora were found to be in-domain and in-genre, even when they were of relatively smaller size.
- **Bilingual Resources.** We make conclusions regarding in-domain and in-genre parallel resources, and out-of-domain and non-parallel resources.
  - **In-domain and In-genre Parallel Resources.** Under this configuration, bilingual-based embeddings were easily the best embedding generation model, outperforming monolingual-based methods for the majority of languages.
  - **Out-of-domain and Non-Parallel Resources.** This setting had more varied results based on the language, the resource, and the vocabulary size. Bilingual-based embeddings still resulted in the best average and overall model, but more languages performed better with monolingual-based methods under this setting.

  - **Out-of-domain Parallel vs. In-domain Comparable.** While the out-of-domain parallel corpus outperformed the comparable corpus overall - 6 languages where comparable does better, and 10 where out-of-domain does better - the comparable corpus does surprisingly well for a considerable number of languages, including languages whose Quran-Bible corpus vocabulary isn't large enough to overcome the domain mismatch.

  The relative size of the target language monolingual, comparable and parallel vocabularies was found to be a factor affecting the performance of bilingual-based vs. monolingual-based embedding generation methods, particularly so for the out-of-domain and non-parallel setting.

- **Bilingual Features.** We make conclusions regarding bilingual feature representations under two settings: in-domain and in-genre parallel resources, and out-of-domain and non-parallel resources.

  - **Features for In-domain Corpora**. The best overall performing feature was found to be our method for combining bilingual sentiment features with target language lexicalization and weight update. We found that training sentiment scores bilingually in this way is more effective than projecting lexicon scores directly from the source language; in addition, pre-trained bilingual sentiment features and lexicalization, when deployed separately, also resulted in some improvements. Bilingual sentiment features were found to help increase the recall of positive and negative sentiment labels and are especially helpful for target languages whose test data is distributed differently for sentiment than the source language. The pre-training corpus was also found to be a factor: it should preferably be evenly distributed for sentiment labels tagged by the lexicon.
  - **Features for Out-of-domain and Comparable Corpora**. More variation among resources was observed in this setting, but bilingual sentiment

features with target language lexicalization was still consistently found to be the best performing feature. In addition, the comparable corpus using only pre-trained sentiment embeddings resulted in improvements for several languages, although not all were significant. Factors affecting performance difference here are the sentiment distribution in the pre-training corpus as well as the target language vocabulary.

Whereas this chapter has assessed the contribution of bilingual representation features and resources to the performance of cross-lingual sentiment models, the next chapter deals with the impact of the choice of source language in the cross-lingual transfer.

Chapter 5

The Role of the Source Language

Thus far in this work, the source language from which the transfer of sentiment occurs has been assumed to be English. This assumption has also been made in the vast majority of current studies involving cross-lingual sentiment analysis, mainly because of the large number of both sentiment analysis resources, such as training datasets or sentiment lexicons, as well as translation resources, such as parallel corpora, that are available for English.

However, the source language can play an important role in the performance of the cross-lingual sentiment model - in particular when the source and target language belong to the same language family or share similar linguistic properties. While the source language may not always be as richly resourced as English, it would be beneficial to understand how the language from which sentiment is transferred affects cross-lingual sentiment performance when equally sized resources are used. Such an analysis would set forth a direction for future research in transferring sentiment from from source languages that are currently moderately-resourced compared to English, such as Arabic or Chinese, and it would faciliate the transfer of sentiment among language families.

In this chapter, therefore, we explore cross-lingual sentiment analysis with a source language other than English, including both Indo-European and non-Indo-European source languages, and with moderately-resourced source languages, such as Arabic, that have until now been only considered as target languages for the purposes of sentiment analysis. In addition to identifying pairs of source and target languages which are best suited for sentiment transfer, our goal is to understand the effect of using a 'pivot' language for the purposes of obtaining a parallel corpus; if a parallel corpus is available between English and Tigrinya, for example, but not between Arabic and Tigrinya, we can use machine translation, which is available for Arabic and English, to obtain a parallel corpus for Arabic and Tigrinya. Finally, we would like to study how preprocessing the source language, particularly in the case of a morphologically rich language such as Arabic, affects the performance of sentiment in the target language. Such studies have been done for machine translation, such as that of Habash and Sadat (2006) for Arabic-English machine translation, but not for cross-lingual sentiment analysis.

Our work in this chapter combines the language family work from our paper (Rasooli et al., 2018) with our more recent work on transferring sentiment from Arabic and Chinese, which we plan to submit for future publication. We start by presenting an experimental analysis of cross-lingual sentiment performance using European and Indo-European source languages, identifying the best source language for each of 17 of our target languages, including English, which under this configuration is considered as a low-resource language with no training data. This analysis, along with the best language pairs, is presented in Section 5.2.

In Section 5.3, we study cross-lingual sentiment analysis with Arabic and Chinese as source languages. This part includes experiments using English as a 'pivot' language to create an Arabic-Tigrinya parallel corpus, and the study of the role of morphological tokenization techniques on the performance of cross-lingual models with Arabic as a source language. Our findings on Arabic preprocessing are consistent with past work on machine translation (Lee, 2004; Habash and Sadat, 2006) which showed that more tokenization and morphological preprocessing helps smaller-sized corpora when translating from Arabic to English, as well as with our own work on Arabic targeted sentiment analysis, which we further detail in future chapters.

Our results across all sets of experiments show that the source language and its properties is an influencing factor in the performance of the cross-lingual model. We find, for example, that European languages in similar sub-families, such as Germanic and Slavic languages, transfer sentiment best from each other, and that transferring sentiment from Arabic to Tigrinya, which is in the same language family, is preferable to doing so from English, even if the Arabic parallel corpus has been machine-translated from English.

Finally, in Section 5.4, we present two new error analyses of our cross-lingual model, now treating English as a target language, and we conclude in Section 5.5.

5.1 Language Families

We briefly re-introduce the language families considered in this chapter. Broadly speaking, our target languages are divided amongst Indo-European (13 languages), Afro-Asiatic (2 languages), Turkic (1 language), Sino-Tibetan (1 language), and Uralic (1 language). Figure 5.1 visualizes these families and their sub-families.

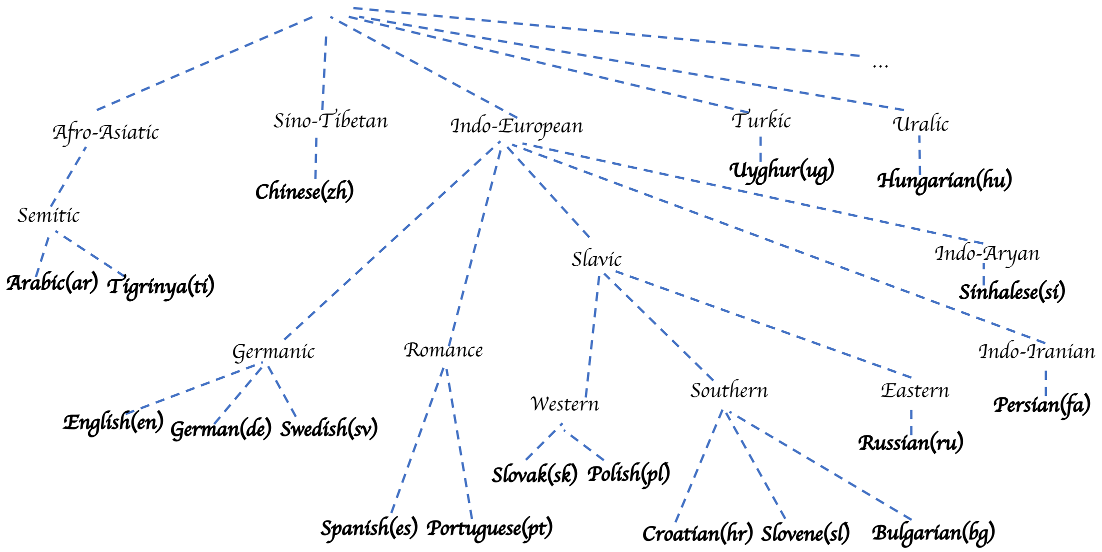


Figure 5.1: Language family tree.

Under the umbrella of Indo-European languages, the largest language family[1] with great variation amongst members, are the Germanic, Romance, and Slavic languages, encompassing Western and Eastern European languages, as well as the Indo-Iranian languages, which include Persian, and the Indo-Aryan languages, which includes Sinhalese.

Sino-Tibetan is the second largest language family, spoken in South, East, and Central Asia, and it includes Chinese.

Afro-Asiatic languages occupy their own branch in the language family tree, with Semitic languages consituting a major sub-branch of this group. Semitic languages originate in the Middle East and include both Arabic and Tigrinya, and this group of families shares some common morphological properties, such as the consonantal root system from which words are formed, and concatenative morphology (e.g attachment of clitics and affixes such as possessive pronouns).

The Uralic language family consists of languages spoken in central and northern Europe and Asia, and includes Hungarian. Turkic, consisting of languages spoken in Eastern Europe and Asia, encompasses Uyghur.

5.2 Transferring Sentiment from European and Indo-European Languages

This section presents our work on cross-lingual sentiment analysis with European and Indo-European source languages. We have grouped these languages together because of their shared properties as well as the large sentiment training datasets that we have been able to acquire for them, with the exception of Sinhalese, which is a low-resource language. In this section, we use European and Indo-European languages as source languages for all other target languages which share a parallel corpus with the

[1] https://www.angmohdan.com/the-root-of-all-human-languages/

source language. Because the Europarl (EP) corpus is multi-parallel - i.e, the same translations are projected across all EP languages - and the Quran-Bible corpus, though not fully multi-parallel, contains translations for almost all source and target language pairs, we are able to use the source language side of these corpora to create bilingual features for transferring sentiment to many target languages.

We describe the experimental setup in Section 5.2.1, and show results in Section 5.2.2.

5.2.1 Experiments

We ran our cross-lingual model architecture, described in Chapter 4. This model has several variations listed in Table 4.1; in what follows, we describe the configuration for this set of experiments. We ran the model using the following source languages, which include both high-resource and moderately-resourced languages:

- **Source Languages:** Bulgarian(bg), English(en), German(de), Spanish(es), Persian(fa), Hungarian(hu), Croatian(hr), Polish(pl), Portuguese(pt), Russian(ru), Slovak(sk), Slovene(sl), and Swedish(sv).

These source-language models were applied for each of the following target languages:

- **Target Languages:** Arabic(ar), Bulgarian(bg), English(en), German(de), Spanish(es), Persian(fa), Hungarian(hu), Croatian(hr), Polish(pl), Portuguese(pt), Russian(ru), Slovak(sk), Slovene(sl), Swedish(sv), Uyghur(ug), and Chinese(zh).

Sinhalese and Tigrinya are excluded from this experiment because of the lack of a shared parallel corpus between these languages and any of the source languages other than English.

5.2.1.1 Bilingual Resources and Features

This set of experiments for European and Indo-European source transfer uses the version of our model which incorporates monolingual-based embedding generation (ML) with Dictionary-Code-Switched embeddings (Dict-CS), target language lexicalization during training (+Lex), Sentiwordnet features (SWN), and cluster embeddings (CL). Experimental configurations for this model are as described in Chapter 4. All source-to-target experiments use this same model, so any difference in results is due to only to the change in the nature and individual configurations of the source language and its resources, and not due to the method for creating bilingual representation features.

The corpora used for creating cross-lingual representation features are the in-domain EuroParl (EP) corpus and the out-of-domain Quran and Bible (QB) corpus. EP translations are available for all our EP languages and QB translations are available for all language pairs except Croatian(hr)-Uyghur, Hungarian(hr)-Ugyhur, Slovak(sk)-Uyghur, and Slovene(sl)-Uyghur. The Uyghur QB corpus contains only Quran translations but no Bible translations (these are not made available as part of the corpus of Christodouloupoulos and Steedman (2014).)

5.2.1.2 Data

We used the untargeted sentiment training and evaluation datasets described in Chapter 3, Section 3.3.1, namely the European Twitter dataset for training and evaluating the European languages, the Persian Product Reviews for training and evaluating Persian, and the untargeted evaluation datasets described for the remaining languages.

All other configurations, including evaluation metric, word embedding method, and text preprocessing, are as described in Chapter 4, Section 4.4.1.

5.2.2 Results

Table 5.1 shows results on identifying sentiment in the target languages using each source language, using the EuroParl corpus, Table 5.2 shows results on identifying sentiment in the target languages using each source language, using the Quran and Bible corpus, and Table 5.3 shows a summary of the best source language for each target language using both parallel corpora.

	Source									
Target	bg	de	en	es	hu	pl	pt	sk	sl	sv
bg	–	42.6	43.5	30.3	36.8	33.1	31.7	33.7	39.3	**44.8**
de	**49.6**	–	45.4	41.7	44.4	46.4	41.5	33.8	43.9	45.9
en	45.4	**49.0**	–	32.5	36.9	47.9	43.7	43.9	46.2	47.9
es	40.8	**41.0**	39.4	–	39.6	40.4	33.3	36.3	36.1	40.8
hu	40.4	40.4	31.8	36.1	–	**48.8**	33.9	45.0	43.1	45.4
pl	47.6	37.2	43.3	24.5	**50.7**	–	34.4	47.5	45.2	46.4
pt	36.7	36.3	39.3	29.6	33.2	35.8	–	31.6	35.7	**39.5**
sk	43.4	39.6	20.4	37.3	32.3	**48.7**	26.0	–	42.0	46.9
sl	**45.7**	33.8	40.1	32.4	36.9	39.9	34.1	37.1	–	39.3
sv	47.1	43.9	**49.0**	29.5	37.8	47.0	36.6	35.2	40.8	–

Table 5.1: Macro-averaged F-measure for predicting cross-lingual sentiment with Indo-European source languages and European Parliament (EP) translation corpus.

5.2.2.1 Discussion

We can quickly notice some patterns across these results: first, that there are languages that tend to transfer well from each other. For example, the Germanic families (English(en), Swedish(sv), and German(de)) transfer well from each other, in addition to being good source languages in general. Using the EuroParl corpus, German is the best source language for English, and English is the best source language for Swedish. With the Quran and Bible corpus, Swedish is the best source language for

	Source												
Target	bg	de	en	es	fa	hr	hu	pl	pt	ru	sk	sl	sv
ar	28.8	28.1	37.3	28.8	21.5	27.8	14.6	33.1	27.3	22.1	**39.1**	27.5	31.9
bg	–	43.1	33.0	34.9	16.1	31.9	30.9	26.5	25.7	41.6	29.8	35.6	**44.3**
de	40.7	–	43.5	32.2	16.3	41.8	36.8	31.9	38.5	44.6	33.7	39.9	**46.5**
en	43.1	47.3	–	35.1	22.3	45.3	33.2	49.5	40.6	48.3	43.0	43.2	**51.7**
es	34.3	36.3	**42.6**	–	33.7	36.1	35.1	33.9	35.4	34.3	37.1	35.5	39.7
fa	28.4	37.3	**40.1**	30.2	–	37.0	33.0	38.5	22.2	26.7	32.3	28.7	29.2
hr	29.7	26.6	30.8	32.7	22.5	–	30.8	33.7	34.7	29.1	**40.4**	37.5	36.8
hu	31.2	34.8	41.1	36.8	29.4	39.1	–	39.4	20.7	**44.4**	40.3	37.7	35.7
pl	**42.2**	40.2	41.7	29.8	26.8	39.7	32.8	–	39.6	36.4	41.4	31.0	39.2
pt	33.3	34.4	38.6	29.6	25.8	**39.1**	28.5	37.3	–	33.6	35.3	33.7	37.9
ru	28.9	39.6	**44.8**	26.5	27.9	38.3	32.7	31.2	33.1	–	32.5	30.0	37.0
sk	17.9	35.1	22.6	26.4	16.4	28.1	35.9	41.1	24.9	31.5	–	32.7	**42.8**
sl	39.0	29.9	32.2	29.5	20.9	**45.5**	31.5	29.3	34.2	34.3	33.5	–	40.5
sv	44.8	**46.7**	39.1	26.1	25.0	33.3	27.4	39.7	32.1	30.6	37.7	31.0	–
ug	25.6	**31.5**	30.0	30.4	24.6	–	–	24.8	15.3	27.8	–	–	27.0
zh	23.7	29.5	30.3	32.8	33.8	30.1	14.9	36.9	22.4	14.7	29.8	21.9	**37.6**

Table 5.2: Macro-averaged F-measure for predicting cross-lingual sentiment with Indo-European source languages and Quran and Bible (QB) translation corpus.

both English and German, and German is the best source language for Swedish.

We observe this pattern with the Slavic languages as well. With the EuroParl corpus and overall, Polish(pl) is easily the best source language for Slovak(sk), its Western Slavic sibling, enabling it to achieve a cross-lingual F-measure of 48.7, much higher than results observed in Chapter 3 when transferring from English. Bulgarian(bg) is the best source language for Slovene(sl), its Southern Slavic sibling, and Polish(pl) transfers well from Bulgarian and Slovak. The Slavic languages also transfer well to and from Hungarian(hu), which is the best source language for Polish (50.7 F- measure). While Hungarian is in its own language family, it does share similarities with Indo-European and in particular Slavic languages. With the Quran and Bible

Target	Best source (QB)	Best source (EP)
Arabic (ar)	39.1 (Slovak)	–
Bulgarian (bg)	44.3 (Swedish)	44.8 (Swedish)
German(de)	46.5 (Swedish)	49.6 (Bulgarian)
English(en)	51.7 (Swedish)	49.0 (German)
Spanish(es)	42.6 (English)	41.0 (German)
Persian(fa)	42.6 (English)	–
Croatian(hr)	40.4 (Slovak)	–
Hungarian(hu)	44.4 (Russian)	48.8 (Polish)
Polish(pl)	42.2 (Bulgarian)	50.7 (Hungarian)
Portuguese(pt)	39.1 (Croatian)	39.5 (Swedish)
Russian(ru)	44.8 (English)	–
Slovak(sk)	42.8 (Swedish)	48.7 (Polish)
Slovene(sl)	45.5 (Croatian)	45.7 (Bulgarian)
Swedish(sv)	46.7 (German)	49.0 (English)
Uyghur(ug)	31.5 (German)	–
Chinese(zh)	37.6 (Swedish)	–

Table 5.3: Macro-averaged F-measure for predicting cross-lingual sentiment with best source languages using European Parliament (EP) and Quran and Bible (QB) translation corpora.

corpus, which includes Croatian(hr), we see that this language is the best language for Slovene(sl), which also belongs to the same sub-family. Similarly, Russian(ru) transfers well to Hungarian and Slovak still does quite well to and from Polish.

Surprisingly, the Romance languages - Portuguese(pt) and Spanish(es) - are not the best source languages for each other, and do better when Germanic source languages are used instead. There must be other factors at play, such as resource sizes in the different source languages; this may be the reason why English is still a better source language for Spanish and Russian - Russian has a larger QB translation corpus with English (454.8K sentences) than with any of the other source languages, as does Spanish (292.6K sentences). However, the EuroParl corpus has the same size for all nine languages, and the Germanic languages are still better sources for Spanish and Portuguese there. The training data for the source languages is similarly sized, so the reasons may be more related to training data quality and sentiment label distribution; for example, Spanish training data is heavily positive-biased and isn't the best source for any target.

For Arabic(ar), we see that Slovak(sk) is the best source language - interesting because Slovak, like Arabic, is highly inflectional and morphologically rich which relsults in a large (i.e, sparse) vocabulary size. Similarly, this may also be a reason why Bulgarian, German, and Swedish transfer well from each other - they have similar vocabulary sizes with the EP corpus, as Table 4.2 in Chapter 4 shows, while the vocabularies of English, Spanish, and Portuguese generated using the same corpus are smaller. For Chinese(zh) and Uyghur(ug), the Germanic languages are the best source languages, for the same reasons mentioned above - larger corpora (English compared to other source languages) and larger vocabulary sizes (German and Swedish compared to English). As with Arabic, a larger vocabulary size for the source language makes it more likely that a semantically and morphologically similar word is recognized in the evaluation data of target language.

Generally speaking, the non-Indo-European languages, namely Arabic, Ugyhur, and Chinese, do not fare as well as the rest when transferring from Indo-European source languages. They are less syntactically and semantically similar to the Indo-European source language families and are thus more likely to incur changes in structure and word ordering when moving from train to test. For these languages, as well as Tigrinya and Sinhalese, running cross-lingual models with European source languages can instead benefit from additional representation features such as lexicalization and bilingual sentiment weights, as shown in Chapter 4. Additionally, the next section looks at using some of these languages as source languages instead.

5.3 Transferring Sentiment from Arabic and Chinese

In this part, we look at transferring sentiment with non-European and non-Indo-European source languages. We study Arabic and Chinese separately because their resource availability - namely parallel corpus sizes and in the case of Arabic, training dataset size - is more limited than that of the languages described in Section 5.2. In order to gain a fair assessment of the contribution of these source languages to the performance of the cross-lingual sentiment models, we therefore configure this set of transfer experiments such that all source languages have equally sized training datasets and parallel data resources, and we additionally sample the datasets such that the sentiment distribution of the training datasets is the same as well.

We consider two approaches, applied to Arabic, in order to further understand the degree towards which the source language makes an impact. The first is the use of an English parallel corpus as a pivot to create a parallel corpus for Arabic and Tigrinya, which are in the same language family. The second is the application of different tokenization methods to preprocess all Arabic text before applying the cross-lingual model. In what follows we describe these approaches, and present experiments and results on Arabic and Chinese transfer in Sections 5.3.3 and 5.3.4.

5.3.1 Pivoting with an English Translation Corpus

The goal of this approach is twofold: to create a parallel corpus that would enable cross-lingual sentiment transfer between Arabic and Tigrinya, and to assess whether an artificially generated machine-translated parallel corpus between two languages in the same language family (Arabic and Tigrinya) performs better or worse than a natural parallel corpus between the target language and a less similar source language (English).

While it is true that a machine translation system is most often not available for a low-resource language like Tigrinya, such a system is available for English and Arabic. We therefore use the LDC English-Tigrinya parallel corpus, described in Chapter 3, and translate the English side of the corpus to Arabic using the Google Translate API[2]. This results in a parallel corpus of the same size for Arabic-Tigrinya. We then use the parallel corpus to generate bilingual-based word embeddings in a shared vector space space for Arabic and Tigrinya, as described in Chapter 4.

5.3.2 Preprocessing and Morphological Richness

The goal of this approach is to identify the effect of preprocessing the source language for cross-lingual transfer, when the source language is morphologically rich. Arabic, for example, exhibits both complex concatenative morphology - how the units of a word join together to form a larger word - as well as derivational morphology - how words can be derived from other words - and inflectional morphology - how words change their form depending on grammatical features. Arabic has eight of these inflectional features: *aspect, mood, person, voice* (applied only to verbs), *case, state* (applied only to nouns and adjectives), *gender* and *number* (applied to both verbs and nominals).

wa+	*sa+*	*y+*	*aktub*	*+uwna*	*+hA*
and	will	3person	write	masculine-plural	it

Table 5.4: Linguistic breakdown of the Arabic word وسيكتبونها.

In addition to inflection, *clitics* can attach to the beginning and end of the inflected base word as follows: `[CONJ+ [PART+ [AL+ BASE + PRON]]]`. Conjugation clitics CONJ (such as *and* +و) come first, followed by preposition clitics PART (such as

[2] https://cloud.google.com/translate/?hl=en

with +ب or *for* +ل), the definite article AL (*the* +ال), followed by the base word, and the pronominal clitics PRON (such as *them* +هم) attach at the end.

Together, these properties mean that a large number of structural and functional variations can exist for any given ‘word’ or lemma, resulting in a rich and often sparse vocabulary. Consider for example the word وسيكتبونها *wasayaktubuwnahA*, ‘and they will write it’ (Table 5.4): the lemma ‘write’ is inflected for 3rd person masculine plural by attaching affixes, and it is also attached to two conjugation clitics and one pronominal clitic.

The work in this part addresses Arabic’s cliticization morphology by applying tokenization techniques. The morphological analyzer MADAMIRA (Pasha et al., 2014) has been trained to split clitics CONJ, BASE, AL, and PRON so that words are broken down into their smaller parts. The tokenization mode ‘D3’ splits off all these clitics (i.e, 3-level decliticization). In the previous chapter, we used the ATB (Arabic Treebank) tokenization, which splits off fewer clitics; these include all types of clitics except the determiner AL, which remains attached. We apply the tokenization to all Arabic text, including parallel corpora and the training dataset, before bilingual feature generation and transfer. The goal is to enable Arabic representation features to become more frequent and less sparse, as well as to reduce the number of out-of-vocabulary words while maintaining the advantage of morphological richness that enables a larger proportion of words in the target language to be represented during source language training.

For a detailed and comprehensive description of the morphological properties of Arabic and their use in NLP, the reader is referred to Habash (2010).

5.3.3 Experiments

We ran our cross-lingual model, described in Chapter 4, and describe feature and resource variations in what follows. We ran the model with Arabic, Chinese, and

English as source languages, and applied it to the following target languages:

- **Target Languages:** Bulgarian(bg), English(en), German(de), Spanish(es), Persian(fa), Hungarian(hu), Croatian(hr), Polish(pl), Portuguese(pt), Russian(ru), Slovak(sk), Slovene(sl), Swedish(sv), Tigrinya(ti) and Uyghur(ug).

In order to control for resource size, we downsampled English and Chinese resources so that they matched the same sizes as that of Arabic, which has the smallest resources of the three languages.

5.3.3.1 Bilingual Resources and Features

This set of experiments for Arabic and Chinese source transfer uses the version of our model which incorporates bilingual-based embedding generation (BL), created directly from parallel corpora using the method of Luong et al. (2015). No added bilingual representation features are included. Experimental configurations for the cross-lingual model are as described in Chapter 4.

The Quran and Bible (QB) corpus was for creating cross-lingual representation features between Arabic, Chinese, and all target languages except Tigrinya, for which we used the LDC Arabic-Tigrinya corpus created as described in Section 5.3.1 instead.

5.3.3.2 Data

We used the untargeted training and evaluation datasets described in Chapter 3, Section 3.3.1, namely the consolidated Arabic training data (Table 3.5), the Chinese Hotel Reviews dataset (Section 3.6), and the evaluation datasets described for the given target languages.

5.3.3.3 Downsampling English and Chinese

Because our English Twitter training dataset (46,622 tweets) and Chinese training dataset (170K hotel reviews) are substantially larger than that of Arabic (8387 tweets), we downsampled each of the English and Chinese datasets to match the same size as the Arabic dataset. Moreover, we sampled the smaller English and Chinese datasets so that they maintained the same distribution of sentiment labels as Arabic (43.5% negative, 22.7% positive, and 22.7% neutral). In this way, any changes in performance of the cross-lingual model are due only to the source language and the content of the training dataset.

In addition, we downsampled the English-to-target and Chinese-to-target Quran and Bible corpora so that the number of parallel sentences used to create bilingual embeddings matched the same size as the Arabic-to-target corpora, which are the smallest of the three languages. Figure 5.2 shows the sizes of the downsampled Quran-Bible corpora for all target languages.

5.3.3.4 Preprocessing Schemes

Before running cross-lingual experiments, we pre-processed all Arabic datasets and corpora with the following two tokenization schemes:

- **ATB**: The Arabic Treebank tokenization method used in Chapter 4, and made available by MADAMIRA.
- **D3**: The 3-level decliticization scheme described in Section 5.3.2 and made available by MADAMIRA.

Table 5.5 shows the vocabulary sizes of source language bilingual embeddings for each target language, with each of the two tokenization schemes for Arabic. We can see that Arabic and Chinese have higher vocabulary sizes than English when using the same parallel corpus (12.7K vocabulary for English vs 17.3K and 15.5K vocabularies

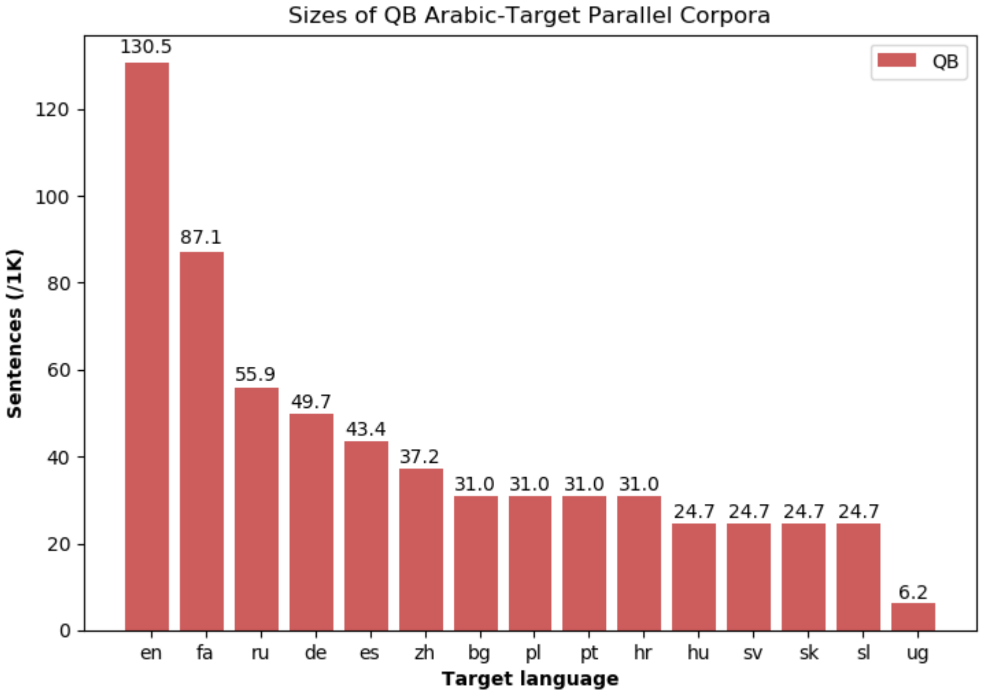


Figure 5.2: Sizes of Arabic-to-target QB parallel data.

for Arabic, and 6.1K vocabulary for English vs 17.4K vocabulary for Chinese), and that the vocabulary size of Arabic is decreased (from 17.3K to 15.5K) by applying D3 tokenization.

5.3.4 Results

Table 5.6 shows the results using the LDC parallel corpus with Tigrinya as a target language, and Table 5.7 shows the results using the QB parallel corpus with all other target languages.

5.3.4.1 Pivoting with a Machine Translated Corpus

From Table 5.6, we can see that the performance of the best cross-lingual model with Arabic as a source language (32.4 F-measure), using the D3 tokenization scheme, is able to outperform the model that uses English as a source language (30.6 F-measure)

	Source Language			
Target Language	en	zh	ar-ATB	ar-D3
ar	12.7	10.3	–	–
bg	5.5	8.8	8.5	7.7
de	7.3	12.3	11.1	10.0
es	6.6	11.2	10.7	9.7
en	–	17.4	17.3	15.5
fa	10.6	23.9	14.3	12.9
hu	4.8	7.0	7.2	6.5
hr	4.8	7.0	7.2	6.5
pl	5.5	8.8	8.5	7.7
pt	5.5	8.8	8.5	7.7
ru	8.2	13.6	14.3	12.9
sk	4.8	7.0	7.2	6.5
sl	4.8	7.0	7.2	6.5
sv	5.5	8.8	8.5	7.7
ti	5.3	–	5.8	5.2
ug	2.4	2.2	1.93	1.86
zh	6.1	–	9.6	8.7

Table 5.5: Vocabulary sizes of source languages for embeddings created from Quran and Bible(QB) and LDC (for Tigrinya) corpora. ‘ar-ATB’ represents the ATB tokenization scheme and ‘ar-D3’ represents the 3-level tokenization scheme. Vocabulary size is represented in 1000 word units.

even when using machine translation to create the Arabic side of the corpus. On the other hand, without this additional tokenization, transferring from Arabic results in a lower score which just exceeds the negative MAJORITY baseline F-measure for Tigrinya, which is 24.0.

This result suggests that with the appropriate processing of the source language, using machine translation between more high-resource source languages would be a beneficial direction to faciliate sentiment transfer towards poorer-resourced languages in the same language family.

	Source		
Target	en	ar-ATB	ar-D3
ti	30.6	24.4	**32.4**

Table 5.6: Macro-averaged F-measure cross-lingual sentiment with English and Arabic using the LDC translation corpus. The Arabic-Tigrinya corpus is machine-translated from the English side to Arabic. The experiment is run 5 times and the averaged result is presented.

5.3.4.2 Effect of Source Languages

Table 5.7 shows that even when using an equal amount of parallel corpora and training data, English still outperforms Arabic and Chinese as a source language for most Indo-European (and some none-Indo-European) target languages.

However, the degree to which this is the case varies by target language, and for a number of target languages, namely Croatian(hr), Slovene(sl), and Slovak(sk), interestingly, transferring from Arabic works better. Not unlike what was observed in Section 5.2.2, languages with larger vocabularies that result from morphological complexity, may make better source-target pairs for transferring sentiment; Slovak, for example was found to be the best Indo-European source language for Arabic. Slovak, Slovene, and Croatian are languages which have larger, more sparse vocabularies, and that may have been why they transferred sentiment better from Arabic.

Considering target languages like Spanish(es), Persian(fa), and Portuguese(pt), which are in the same language family as English, the results of using Arabic or English as source languages are quite close, and one explanation for this could be the historical borrowing of vocabulary from these languages and Arabic. Considering transferring to Chinese, English and Arabic (with the best model) do equally well as source languages; this would make sense as the three languages are all in completely separate language families. For transferring to Uyghur, however, English does substantially better as a source language than either Arabic or Chinese, which is somewhat surprising given that the Uyghur language has been influenced by both

	Source			
Target	en	ar-ATB	ar-D3	zh
ar	**37.5**	–	–	29.8
bg	**44.7**	39.2	39.7	31.7
de	**43.5**	35.2	35.7	31.6
en	–	**43.8**	42.8	39.6
es	**39.5**	38.0	38.7	31.7
fa	**49.7**	48.5	48.3	47.2
hu	**44.6**	34.3	33.7	30.4
hr	38.5	**41.5**	40.8	38.8
pl	**39.2**	34.1	34.4	33.7
pt	**40.6**	38.4	37.8	32.7
ru	**45.7**	31.1	33.4	38.1
sk	35.6	36.6	**40.9**	32.1
sl	36.7	33.9	**37.1**	34.5
sv	**41.6**	37.1	36.8	34.2
ug	**38.6**	32.5	29.8	21.2
zh	**42.7**	34.9	42.6	–
AVG	**41.2**	37.3	38.2	33.8

Table 5.7: Macro-averaged F-measure for predicting cross-lingual sentiment with English(en), Chinese(zh), and Arabic(ar) using the QB translation translation corpus. The experiments are run 5 times and the averaged result is presented.

Arabic and Chinese. It is possible that the small Uyghur QB corpus with resulting 2K vocabulary size is too small to have effected positive learning of bilingual representational features.

It is unsuprising that English is easily the best source language for German(de) and Swedish(sv), and it is also so for Russian(ru) and Bulgarian(bg). Arabic achieves higher F-measures in transferring to target languages than does Chinese, but these results could have been influenced by the genre of the training data, which is Twitter for both Arabic and English but hotel reviews for Chinese, and therefore no strong conclusion can be drawn here.

On a final note, we can see that using the same training data size, genre, and parallel corpus size, transferring sentiment from Arabic to English yields a higher score (43.8) than transferring sentiment from English to Arabic (37.5). This is consistent with the task of machine translation into morphologically complex languages, where typically BLEU (Papineni et al., 2002) scores for Arabic-English machine translation are higher than BLEU scores for English-Arabic machine translation.

5.3.4.3 Effect of Preprocessing

On average, preprocessing Arabic with a tokenization scheme that uses morphological disambiguation to separate all types of clitics positively affects the transfer of sentiment from Arabic into other target languages by reducing vocabularity sparsity. This is clearly the case with Tigrinya(ti), Slovak(sk), Slovene(sl), and Chinese(zh), but not so for other languages, like English, and makes virtually no difference for Persian(fa), German(de), and Bulgarian(bg). It was shown by Habash and Sadat (2006) that full decliticization schemes work especially well for machine translation when using small-sized parallel corpora; this is likely a factor here as Persian and German have relatively larger QB corpora while Slovak and Slovene have smaller ones (Figure 5.2).

5.4 Error Analysis

We present two error analyses using the output of the cross-lingual model with English as a target language. In the first, we use European and Indo-European source languages, and apply a new ensemble that consists of combining the mixed-language training data of *all* source languages described in Section 5.2, and training a single cross-lingual model with multilingual code-switched DICT-CS embeddings using EP corpus supervision. The ensemble model was presented in our group paper (Rasooli et al., 2018) and results in improvements for several languages when combining data from multiple source languages. This model is applied to English as a target language and results in an F-measure of 54.0, topping the best F-measure of 51.7 obtained when transferring from Swedish.

In the second analysis, we examine the output of the best model trained on Arabic and applied to English, which resulted in an F-measure of 43.8 using ATB tokenization and 42.8 using D3 tokenization.

5.4.1 English as a Target Language with European and Indo-European Sources

This error analysis was conducted in order to better understand the kinds of errors made by the cross-lingual model and whether they result from the deep learning model itself or from the transfer to a different language. We sampled 66 errors at random from the output of the cross-lingual model trained on European and Indo-European sources, and compared its predictions with both the gold labels and with a supervised model trained on English.

Generally, we found that the source of sentiment errors comes from the following reasons (Table 5.8): a key sentiment indicator was missed (e.g., “love,” “excited”, “bored”), there were misleading sentiment words (e.g., “super” in context of “getting up super early”, “handsome” in context of a question), the tweet contained misspelled/rare words (e.g., “bff,”, “bae”, “puta”), inference was required (e.g., “i need to seriously come raid your closet” is positive without containing positive words), the correct answer was not clear or not easily determined for a human annotator (e.g., “a mother’s job is forever”), or the gold label was clearly wrong (e.g “thanks for joining us tonight! we kept it as spoiler free as possible!” has a neutral instead of positive gold label). There are thus many tweets in this error sample where the sentiment is not clear cut.

Error Type	Example
Sentiment indicator missed	rt **bored** of my chilled weekend already *(predicts positive; gold negative)*
Misleading sentiment words	up **super** early to have my boy at his ffa judging comp *(predicts positive; gold negative)*
Mispelled or rare words	rt awwwww, **imbecil** . *(predicts positive; gold negative)*
Inference required	walking socks take up so much space ! *(predicts neutral; gold negative)*
Gold wrong	i filled out ova 30 job applications *(predicts neutral; gold positive)*

Table 5.8: Errors made by the European and Indo-European cross-lingual model when transferring to English.

To study the kinds of errors resulting from the language transfer as opposed to the machine learning model itself, we divided the error samples into four groups:

1. In the first group (48.5% of cases), the supervised model makes a correct prediction, but the cross-lingual model results in an error. Looking at examples in this group, we found that this often occurs when the English target data contains rare, mispelled, or informal language words which are unlikely to have been learned using cross-lingual representations from a parallel corpus such as EuroParl.

    - *"**fck na** ! ! marshall ! bear nation hopes your **aight** ! ! !"* (negative, transfer predicts positive)
    - *"eagles might get **doored** tonight :'("* (negative, transfer predicts positive)

2. In the second group (26% of cases), the supervised model and the cross-lingual model make the same error and thus the cause for the error likely comes from the model rather than the transfer. We determined that 6 of these cases have

an incorrect gold label, and 11 result from errors of the supervised model where the answer was unclear, key sentiment was missed, or inference was required.

- *“don ’t let anyone discourage you from following your dreams ! it was one of the best decisions i made because it changed my lif ...”* (gold negative [wrong], transfer and supervised predict positive)
- *“can ’t wait to be an uncle again a wee boy this time , surely his names got to be jack if no , at least make it his middle name”* (gold positive [requires inference], transfer and supervised predict neutral)

3. In the third group (16.6% of cases), the supervised and cross-lingual models make different kinds of errors and thus the source of the error is likely from both the model itself and the transfer. We determined that three of these cases have an incorrect gold label, and the remaining eight are an error of the supervised model where the answer was unclear, key sentiment was missed, inference was required, or the sentence contained misleading sentiment words.

   - *“mount gambier that was rad and sweaty as hell , just one show left on the tour for us tomorrow in adelaide ”* (gold positive [misleading sentiment], supervised neutral, transfer negative)

4. In the fourth group (9% of cases), the gold and supervised models agree, but the cross-lingual model, which was trained on different data, actually makes a better prediction.

   - *“this photo taken on 9th september with high quality one of my bday gifts from my friend thank you brother”*(gold and supervised predict neutral, transfer predicts positive)

About half of the errors are clearly because of the transfer to a different language, but there are also a good number of cases where even the supervised model makes the same error as the transfer model. The errors that the cross-lingual model makes are reasonable because of the peculiarities and difficulties of the language of our Twitter evaluation data. In future studies, a comparable corpus could be collected by scraping Twitter in a manner to which our Wikipedia corpus was collected, and bilingual BL or DICT-CS representations of Twitter-specific vocabulary could be learned from there.

5.4.2 English as a Target Language with Arabic as a Source

We compared the output of the Arabic-to-English transfer model (under D3 tokenization) with a supervised English model trained on the same downsampled and sentiment distributed training dataset. This supervised model results in an accuracy of 62.3, macro-averaged F-measure of 61.4, positive F-measure of 58.1, negative F-measure of 59.1, and neutral F-measure of 66.9. In contrast, the transfer model results in an accuracy of 43.2, macro-averaged F-measure of 42.8, positive F-measure of 43.3, negative F-measure of 38.2, and neutral F-measure of 46.8.

We sampled and analyzed 60 errors from these two models. We found similar categories of errors as with transferring from European and Indo-European languages to English - namely, those shown in Table 5.8 - however when transferring from Arabic, we observed more of errors like 'sentiment indicator missed' compared to the European and Indo-European model, where more errors were due to mispellings, misleading sentiment words, and requiring inference. Additionally, because of the small size of the training data and the negative bias in the distribution of sentiment, we observed many errors where the model predicted 'negative' sentiment due to majority baseline influence, even though the tweet contained no negative sentiment indicators.

We again divided the error samples into four groups:

- In the first group (51.7% of cases), the supervised model makes a correct prediction but the cross-lingual model result in an error. The majority of errors in this group come from a negative majority baseline influence or a missed key sentiment indicator.

  - *do you know what you wanna do when you're done w/ school yet?* (neutral, transfer predicts negative)
  - *rt i have a crush on fall weather, hot drinks, and cozy sweaters* (positive, transfer predicts negative)

- In the second group (23.3% of cases), the supervised model and the cross-lingual model make the same error. These errors mostly required inference or came from a wrong gold annotation.
  - *please mention me i really want to reach my goal x37* (gold neutral[wrong or unclear], transfer and supervised predict positive)
  - *for my birthday i got a humidifier and a de-humidifier ... i put them in the same room and let them fight it out* (gold positive[requires inference], transfer and supervised predict negative)

- In the third group (15% of cases), the supervised and cross-lingual models make different kinds of errors. These again were due to a variety of causes, such as wrong gold, misleading sentiment words, missing a key sentiment indicator, or requiring influence.
  - *rt 1 more day until this is back im screaming* (gold positive[requires inference], transfer predicts neutral, supervised predicts negative)

- In the last group (10% of cases), the gold and supervised models agree, but the cross-lingual model actually makes a better prediction.
  - *marriott hotels servers up a " fresh " approach - healthy vending machine debuts* (gold neutral[wrong], transfer predicts positive, supervised predicts neutral)

The distribution of groups and output of the cross-lingual model relative to the supervised model is more or less consistent with that observed when transferring from

European and Indo-European languages.

5.5 Conclusion

This chapter studied the influence of the source language and its characteristics when transferring sentiment cross-lingually. In contrast to most previous work which assumes that the source language is English, we evaluated the performance of cross-lingual sentiment models when trained on European and Indo-European languages, as well as Arabic and Chinese. Moreover, to facilitate the transfer of sentiment from Arabic, we introduced new techniques such as pivoting with machine translation to create an Arabic-Tigrinya corpus, and applying preprocessing schemes to reduce the sparsity of bilingual features that arise from morphological complexity. Our findings, summarized below, point to the important role played by the source language when transferring sentiment cross-lingually and the need for a future direction towards increasing resources made available to moderately resourced languages such as Slovak, Arabic, or Chinese, to faciliate transfer to target languages in similar language families.

- **Language families**: Languages from similar language families transfer sentiment well from each other. This was especially the case for the Germanic and Slavic languages, and evident in the performance of English compared to Arabic and Chinese when transferring to most Indo-European languages, even when using similarly sized resources. The success of language family transfer for sentiment analysis is consistent with past results on other cross-lingual tasks, such as direct transfer of part-of-speech tagging (Kim et al., 2017).
- **Resource sizes and distribution**: Languages with large parallel resources and evenly distributed sentiment datasets are generally good source languages,

as demonstrated by the success of languages like English (large parallel corpus) and Swedish (balanced dataset) when transferring to other European languages.

- **Morphological richness**: Languages with similar morphological complexity and vocabulary sizes transfer sentiment well from each other. This is demonstrated by the success of sentiment transfer amongst languages like German, Bulgarian, and Swedish, or Arabic, Slovak, Croatian, and Tigrinya, which are similar in vocabulary size. Moreover, applying high-resource morphological tokenization schemes enables Arabic to transfer sentiment better on average and is consistent with past results on machine translation.

Our error analysis with English as a target language revealed that Twitter-specific out-of-vocabulary words, which are unlikely to occur in a translation corpus or Wikipedia comparable corpus, are a source of error in the model; future work for improving the performance of untargeted cross-lingual sentiment models could thus focus on the collection and learning of bilingual embeddings from Twitter and social media corpora. In the next chapter, we turn to targeted sentiment analysis, where we focus on identifying sentiment towards targets in short documents.

Chapter 6

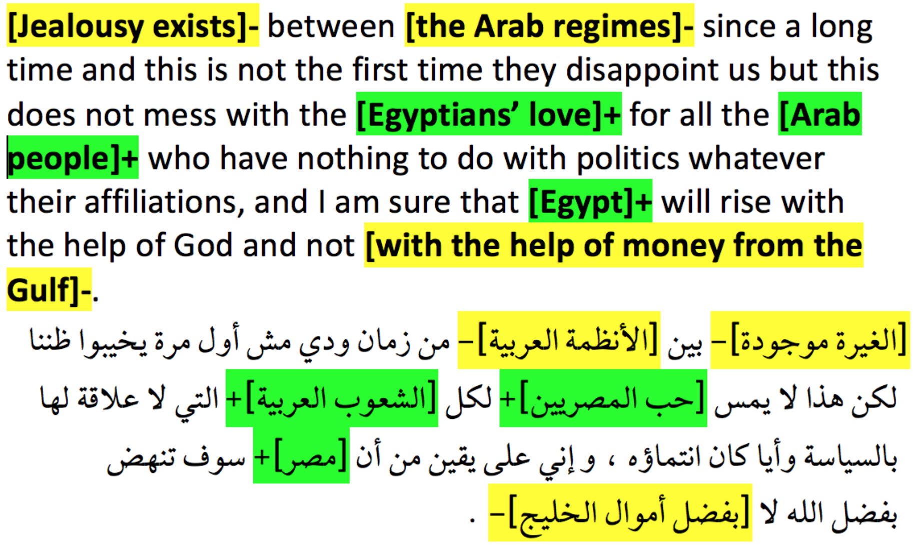
Targeted Sentiment Rooted in Documents

The expression of sentiment in language often does not occur in an isolated context, but is instead usually directed towards a topic, such as an entity, event, issue or a situation - a target of sentiment. Knowledge of the target is important for making sense of the sentiment expressed; consider, for example, the following text:

Example 6.1. `The` **`will of the people`** `will prevail over the` **`regime's brutality.`**

The sentiment expressed here by the text is positive towards 'the people' but negative towards 'the regime'. A model that can identify the sentiment expressed towards specific targets is therefore more informative than one which only identifies the overall sentiment of the text.

Targeted sentiment analysis has been studied extensively in natural language processing, but it has usually focused on English (with some studies in other languages, such as that of Al-Smadi et al. (2015)), and more often than not it has focused on named entity targets, or targets that have already been specified in the text (Jiang et al., 2011; Biyani et al., 2015). The targeted sentiment problems addressed in this chapter cover long and often complex spans of text that may contain multiple entities or events, and they are not restricted to named entities, as shown in Figure 6.1. Moreover, the target of sentiment and the segment of text expressing sentiment towards the target need not always occur in the same sentence, necessitating global methods for associating the two. In some cases, the target of the sentiment need not even be an entity that is mentioned explicitly in the text, but can instead constitute



[Jealousy exists]- between **[the Arab regimes]-** since a long time and this is not the first time they disappoint us but this does not mess with the **[Egyptians' love]+** for all the **[Arab people]+** who have nothing to do with politics whatever their affiliations, and I am sure that **[Egypt]+** will rise with the help of God and not **[with the help of money from the Gulf]-**.

[الغيرة موجودة]- بين [الأنظمة العربية]- من زمان ودي مش أول مرة يخيبوا ظننا
لكن هذا لا يمس [حب المصريين]+ لكل [الشعوب العربية]+ التي لا علاقة لها
بالسياسة وأيا كان انتماؤه ، و إني على يقين من أن [مصر]+ سوف تنهض
بفضل الله لا [بفضل أموال الخليج]- .

Figure 6.1: Arabic text and English translation with multiple annotated target entities and sentiment (green:pos, yellow:neg).

a higher-level 'situation' or category which itself can encompass multiple entities or events. These kinds of problems fall into the vein of targeted sentiment analysis that is rooted in short documents, sharing similarities with problems such as stance detection (Somasundaran and Wiebe, 2009; Mohammad et al., 2016a) using sentiment targets to identify stance (Farra et al., 2015b), or fine-grained sentiment analysis systems aimed at predicting entity and event-level targets as well as sources and the polarity of sentiment (Deng and Wiebe, 2015a).

Targeted sentiment rooted in documents shares some similarities with the sentiment analysis task of predicting consumer sentiment in customer reviews along with their *aspects* (e.g 'service' of a restaurant, or 'speed' of a laptop) (Hu and Liu, 2004; Pontiki et al., 2014). However, aspects are more limited as targets of sentiment, while the texts considered in this chapter are more open in domain and therefore pose a greater challenge for sentiment identification; customer review datasets are usually focused on a single product, such as 'restaurants' or 'laptops', while the documents in this chapter may span multiple entities or events and are not restricted to a single domain.

Targeted Sentiment in Documents	
Open-Domain Dataset Annotation for Arabic	Section 6.1
Open-Domain Target and Sentiment Identification Models	Section 6.2
Identification of Sentiment towards Situation Frames	Section 6.3

Table 6.1: Roadmap of chapter on targeted sentiment rooted in documents.

The motivation for studying these tasks in the midst of our larger cross-lingual and low-resource goals is twofold: first, to study document-rooted and open-domain targeted tasks in a moderately-resourced language (Arabic), identifying the characteristics of the language that affect the performance of targeted sentiment, and second, to explore and introduce even more complex problems, such as the task of identifying sentiment towards situations, both with the goal of enabling further research in cross-lingual transfer of targeted sentiment, a topic we introduce in the last chapter of the thesis.

We thus consider the problem of annotating as well as identifying open-domain targeted sentiment in short Arabic documents, before proceeding to the situation frame task, where we introduce and briefly study a new problem: that of identifying sentiment towards situations in English and Spanish.

In Sections 6.1 and 6.2, we describe our work on open-domain targeted sentiment in Arabic. We present a new dataset of news article comments that we collected for this problem (Farra et al., 2015a), and develop an approach for identifying important entities along with their sentiment in Arabic documents (Farra and McKeown, 2017). Both the dataset and our code[1][2][3] are publicly available. Through our analysis, we demonstrate the impact of segmentation techniques on the identification of both targets and the sentiment towards them in Arabic, and find, as we did in Chapter

[1] https://www.cs.columbia.edu/˜noura/Resources.html

[2] https://github.com/narnoura/SentimentTargets-paper/tree/master/data/arabic-finegrained

[3] https://github.com/narnoura/SentimentTargets-paper

5, that the morphology of the language plays an important role in the analysis of sentiment. In Section 6.3, we introduce the problem of identifying sentiment towards situations, present preliminary results, and suggest directions for future research.

6.1 Collecting an Arabic Open-Domain Targeted Dataset

Annotating targets of opinion is a difficult and expensive task, requiring definition of what constitutes a target, whether targets are linked to sentiment expressions, and how the text spans of targets should be defined (e.g **'the people'** vs. **'the will of the people'** or **'the regime'** vs. **'the regime's brutality'**), a problem which annotators often disagree on (Pontiki et al., 2014; Kim and Hovy, 2006; Somasundaran et al., 2008).

Additionally, it is not always straightforward to attribute a target to a specific sentiment expression, as some annotation schemes have proposed. Consider for example the following text:

Example 6.2. `The Lebanese Member of Parliament said he was convinced that there would be a consensus on` **`the presidential election`**`, because since the moment the United States and Iran had reached an understanding in the region, things were starting to look positive.`

It is not clear that there is a single sentiment expression that leads us to believe that the Member of Parliament is optimistic about the target **presidential election**; it could be 'convinced', 'consensus', 'reached an understanding', 'look positive', or a combination of the above. Such decisions are difficult for annotators to agree on; many studies have noted these challenges (Stoyanov and Cardie, 2008; Ruppenhofer et al., 2008) which can make the task complex.

Compared to the amount of resources available for sentiment analysis, there is much less annotated data available for this more fine-grained type of analysis, even for high-resource languages. Due to the difficulty of the task, most of the available datasets of fine-grained sentiment analysis have been annotated by trained annotators or expert linguists, making the process slower and more expensive. This makes the problem of transferring targeted sentiment even more significant for low-resource languages, as will be discussed in Chapter 7.

The work described in this section considers annotation of targets using a sequence of simple crowdsourced sub-steps. We focus on Arabic, where there are much fewer publicly available resources for targeted sentiment analysis, and where concatenative morphology proposes an interesting challenge for defining target entity spans. We assume that any nominal phrase can be a target of sentiment: people, places, events, or concepts, and we develop a two-stage annotation process for annotating targets using the crowdsourcing platform Amazon Mechanical Turk[4]. In the first stage, annotators list all important noun phrase entities, and in the second stage, they choose the polarity expressed (positive, negative, or neutral) towards any given entity. We select online data from multiple domains: politics, sports, and culture; and we provide a new publicly available resource for Arabic by annotating it for targets of opinions along with their polarities. Finally, we evaluate the quality of the data at different stages, obtaining majority agreement on sentiment polarity for 91.8% of entities in a corpus of 1177 news article comments. Section 6.1.1 describes the annotation process, Section 6.1.3 describes how we selected the data for annotation, and Section 6.1.4 presents an analysis of the targeted dataset.

[4] https://www.mturk.com/

6.1.1 Annotation Process

We assume targets of opinions to be nominals representing entities, events, or concepts; for example, targets can include politicians, organizations, events, sports teams, companies, products, concepts such as ‘democracy’, or entities representing ideological belief.

Example 6.3. `It is great that so many people showed up to` **`the protest.`**

In the above example, the full target span is the clausal phrase ‘that so many people showed up to the protest’, representing the object of ‘great’. However, the actual entity which receives the positive sentiment is **‘the protest’**. We are interested in annotating such entities, as this would enable the development of a targeted model that could ‘summarize’ sentiment towards different entities in the short document.

Given the complexity of the task, we annotate targets without specifying the specific sentiment expressions that are linked to them, as in Pontiki et al. (2014); Hu and Liu (2004), although the dataset can be extended for this purpose to provide richer information for modeling. We don’t consider targets of subjective-neutral judgments (e.g *‘I expect it will rain tomorrow’*). For this corpus, as with the rest of the thesis, we are interested only in targets of polar positive or negative sentiment; all other text is regarded as neutral. Finally, since our data comes from comments to online newspaper articles, it is assumed that the source of the expressed sentiment is the writer of the post, although this does not affect the identification or annotation of targets.

6.1.2 Amazon Mechanical Turk Tasks

Instead of asking annotators to directly identify targets of opinions, which we believed to be a more difficult task, we broke the annotation into two stages, visualized in

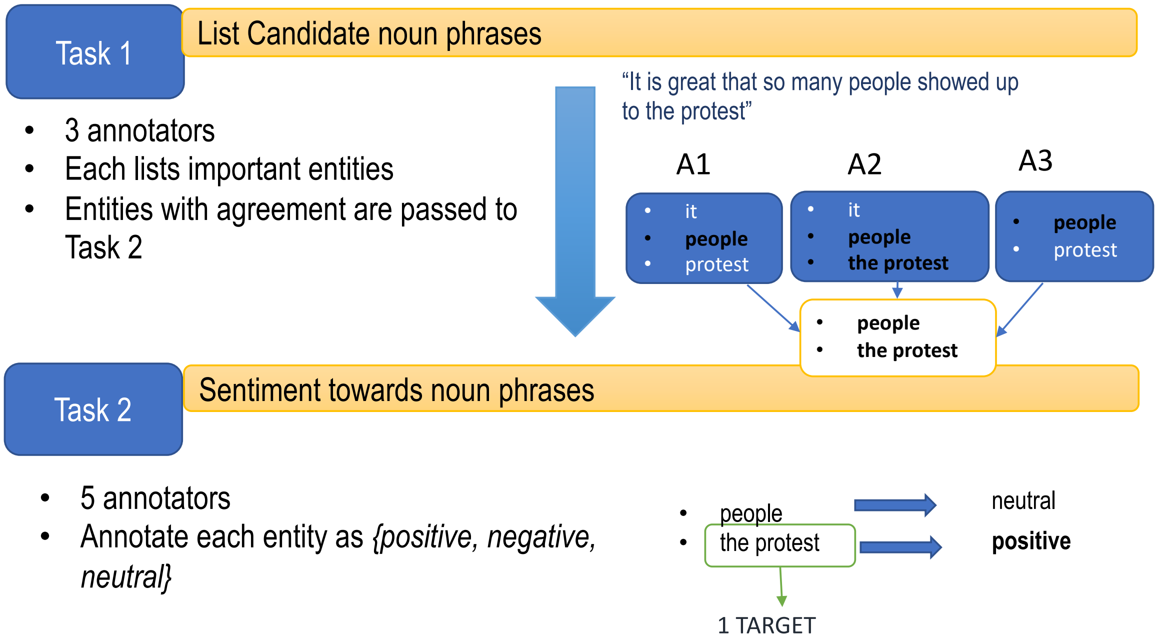

Task 1
List Candidate noun phrases
3 annotators
Each lists important entities
Entities with agreement are passed to Task 2
"It is great that so many people showed up to the protest"
A1
A2
A3
it
people
protest
it
people
the protest
people
protest
people
the protest
Task 2
Sentiment towards noun phrases
5 annotators
Annotate each entity as {positive, negative, neutral}
people
the protest
neutral
positive
1 TARGET

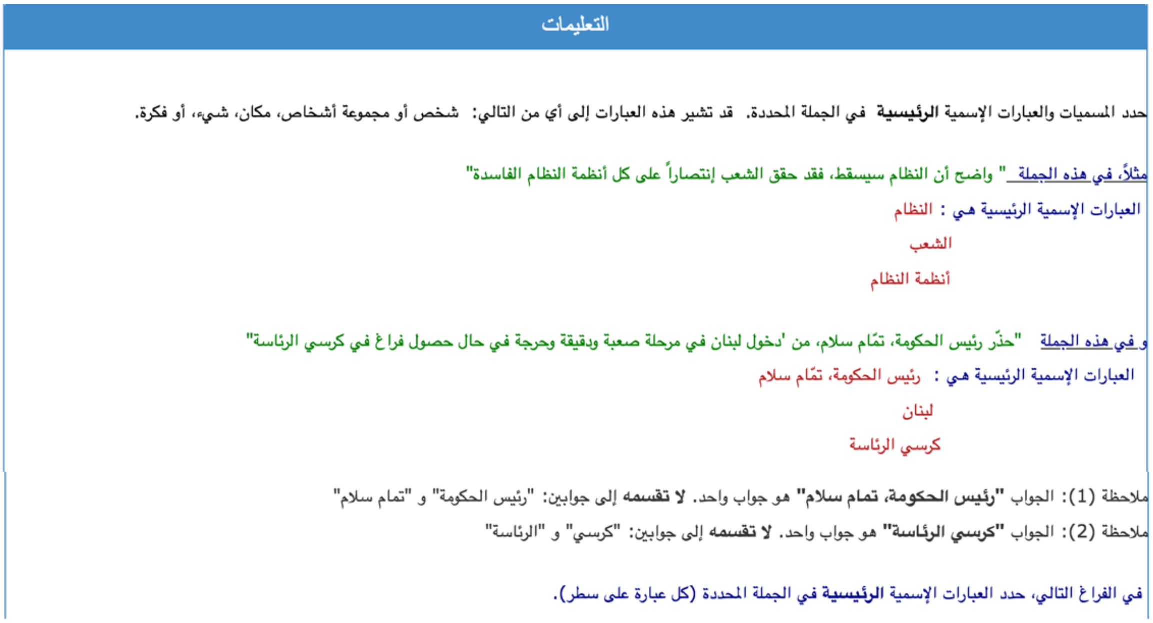
Figure 6.2: Annotation process for Arabic open-domain targets of sentiment.

Figure 6.2, each in a different series of HITs (Human Intelligence Tasks). The task guidelines were presented in Modern Standard Arabic (MSA) to guarantee that only Arabic speakers would be able to understand and work on them. Many of the insights in the task design were gained from an extensive pilot study.

6.1.2.1 Task 1: Identifying Candidate Entities

Given an article comment, annotators were asked to list the main nouns and noun phrases that correspond to people, places, things, and ideas. This task, or HIT, was given to three annotators and examples of appropriate answers were provided. A sample screenshot is provided in Figure 6.3.

The answers from the three annotators were then combined by taking the intersection of common noun phrases listed by all three responses. If annotators only agreed on a subset of the noun phrase, we chose the maximal phrase among agreed entities in order to determine the entity span. For example, if two annotators specified 'the president' and a third specified 'the election of the president', we



التعليمات

حدد المسميات والعبارات الإسمية **الرئيسية** في الجملة المحددة. قد تشير هذه العبارات إلى أي من التالي: شخص أو مجموعة أشخاص، مكان، شيء، أو فكرة.

مثلاً، في هذه الجملة " واضح أن النظام سيسقط، فقد حقق الشعب إنتصاراً على كل أنظمة النظام الفاسدة"

العبارات الإسمية الرئيسية هي : النظام
الشعب
أنظمة النظام

و في هذه الجملة "حذّر رئيس الحكومة، تمّام سلام، من 'دخول لبنان في مرحلة صعبة ودقيقة وحرجة في حال حصول فراغ في كرسي الرئاسة"

العبارات الإسمية الرئيسية هي : رئيس الحكومة، تمّام سلام
لبنان
كرسي الرئاسة

ملاحظة (1): الجواب **"رئيس الحكومة، تمام سلام"** هو جواب واحد. **لا تقسمه** إلى جوابين: "رئيس الحكومة" و "تمام سلام"
ملاحظة (2): الجواب **"كرسي الرئاسة"** هو جواب واحد. **لا تقسمه** إلى جوابين: "كرسي" و "الرئاسة"

في الفراغ التالي، حدد العبارات الإسمية **الرئيسية** في الجملة المحددة (كل عبارة على سطر).

Figure 6.3: Screenshot of instructions for task 1 HIT.

kept ‘the election of the president’. The maximal noun phrase was also chosen by Pontiki et al. (2014) when resolving disagreements on target spans. We allowed annotators to list references in the comment to the same entity (e.g ‘The president’ and ‘President Mubarak’) as separate entries.

Insights from Pilot. We asked specifically for the ‘main’ noun phrases, after we found that annotators in the pilot over-generated nouns and noun phrases, listing clearly unimportant entities (such as اليوم *‘today/this day’*, and السلام *‘hello/the greeting’*), which would make Task 2 unnecessarily expensive. They would also break up noun phrases which clearly referred to a single entity (such as separating كرسي *‘the seat’* and الرئاسة *‘the presidency’* from كرسي الرئاسة *‘the presidency’s seat’*), so we instructed them to keep such cases as a single entity. These reasons also support choosing the maximal agreeing noun phrase provided by annotators. By making these changes, the average number of entities resolved per comment was reduced from 8 entities in the pilot study to 6 entities in the full study.

We paid 30 cents for Task 1, due to the time it took workers to complete (2-3 minutes on average).

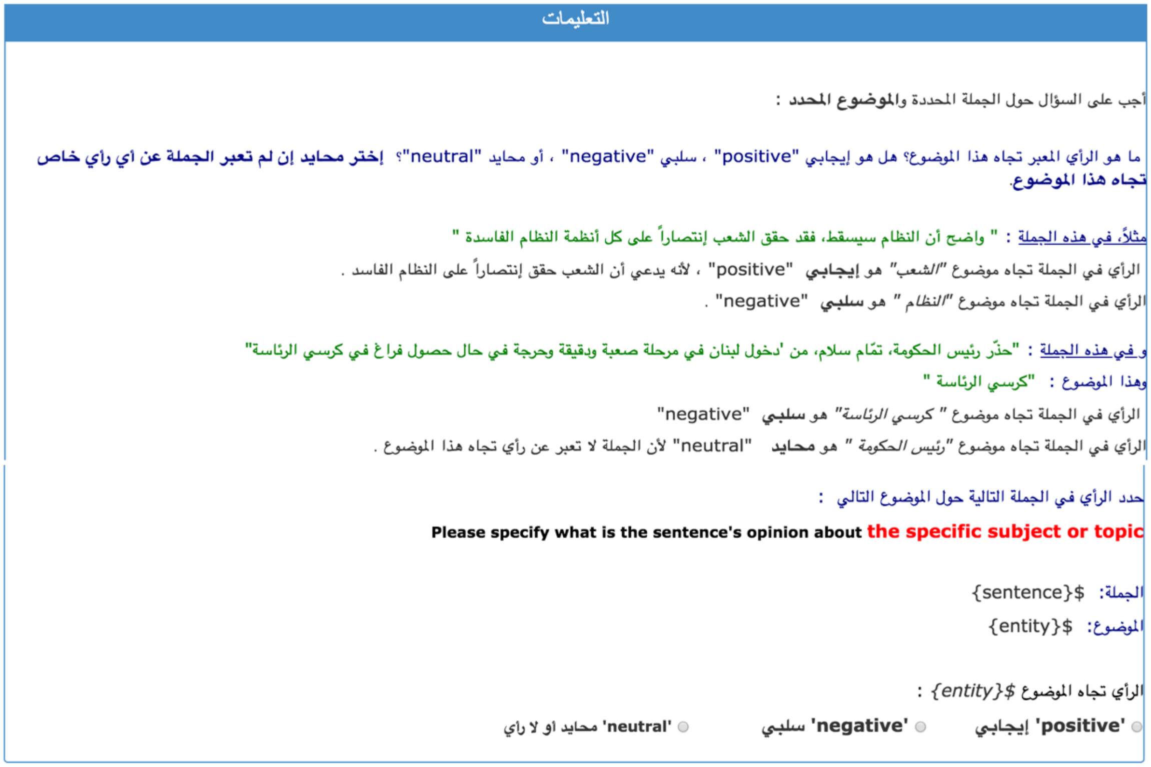
6.1.2.2 Task 2: Identifying Sentiment towards Entities

In the second task (HIT), annotators were presented with an article comment and a single entity, and were asked to specify the opinion of the comment towards the given 'topic'. The entities were chosen from the resolved responses in Task 1. The question was presented in multiple-choice form where annotators could choose from options 'positive', 'negative', or 'neutral'. Each HIT was given to five annotators, and the entities resolved to 'positive' or 'negative' with majority agreement are considered to be targets of sentiment. Entities with neutral majority are discarded as non-targets, while entities with disagreement on polarity (e.g two annotators assign negative sentiment while one assigns positive sentiment) are kept aside for future use as will be described in Section 6.2.

In this question, we told annotators that sentiment could include opinions, belief, feelings, or judgments, and that the 'neutral' option should be selected if the text reveals either no sentiment or an objective opinion towards this particular entity. We provided multiple examples. For this task, we paid workers 5 cents per HIT, which took 30 seconds to 1 minute to complete on average.

Insights from Pilot. In our pilot study, we had an additional question in this HIT which asks annotators to specify the source of the sentiment expression, which could be the writer or someone else mentioned in the text. However, we removed this question in the final study due to the low quality of responses in the pilot, some of which reflected misunderstanding of the question or were left blank.

Additionally, we found that some annotators specified the overall sentiment of the comment rather than the sentiment about the topic. We thus emphasized, and



التعليمات

أجب على السؤال حول الجملة المحددة و**الموضوع المحدد** :

ما هو الرأي المعبر تجاه هذا الموضوع؟ هل هو إيجابي "positive" ، سلبي "negative" ، أو محايد "neutral"؟ **إختر محايد إن لم تعبر الجملة عن أي رأي خاص تجاه هذا الموضوع**.

مثلاً، في هذه الجملة : " واضح أن النظام سيسقط، فقد حقق الشعب إنتصاراً على كل أنظمة النظام الفاسدة "

الرأي في الجملة تجاه موضوع *"الشعب"* هو **إيجابي** "positive" ، لأنه يدعي أن الشعب حقق إنتصاراً على النظام الفاسد .

الرأي في الجملة تجاه موضوع *"النظام "* هو **سلبي** "negative" .

و في هذه الجملة : "حذّر رئيس الحكومة، تمّام سلام، من 'دخول لبنان في مرحلة صعبة ودقيقة وحرجة في حال حصول فراغ في كرسي الرئاسة"

وهذا الموضوع : "كرسي الرئاسة "

الرأي في الجملة تجاه موضوع *" كرسي الرئاسة"* هو **سلبي** "negative"

الرأي في الجملة تجاه موضوع *"رئيس الحكومة "* هو **محايد** "neutral" لأن الجملة لا تعبر عن رأي تجاه هذا الموضوع .

حدد الرأي في الجملة التالية حول الموضوع التالي :

Please specify what is the sentence's opinion about the specific subject or topic

الجملة: {sentence}$

الموضوع: {entity}$

الرأي تجاه الموضوع *{entity}$* :

○ **'positive'** إيجابي ○ **'negative'** سلبي ○ **'neutral'** محايد أو لا راي

Figure 6.4: Screenshot of instructions for task 2 HIT.

included an additional English translation of the instruction that the opinion polarity should be about the specific topic and not of the whole comment. A sample screenshot is shown in Figure 6.4.

We completed the full annotation study in five rounds of a few hundred comments each. For the first two rounds of annotation, we rejected all HITs that were clearly spamming the task or were not Arabic speakers. After that we created task qualifications and allowed only a qualified group of workers (5 for Task 1 and 10 for Task 2) to access the tasks, based on their performance in the previous tasks.

6.1.3 Data Selection

The annotation data was selected from the Qatar Arabic Language Bank (QALB) (Mohit et al., 2014; Zaghouani et al., 2014), which includes online comments to Al-

Domain	# Comments	Distribution(%)
Politics	596	51
Culture	382	32
Sports	199	17
Total	1177	100

Table 6.2: Distribution of selected article comments by domain.

jazeera[5] newspaper articles.

6.1.3.1 Topic Modeling

We initially selected a random sample of data from the Aljazeera corpus, the majority of which comes from the politics domain. In the pilot study and first annotation round, we found that this data was biased towards negative sentiment. We thus used topic modeling (Blei et al., 2003), with the mallet toolkit implementation (McCallum, 2002), to select data from other domains which were more likely to contain positive expressions of sentiment. Upon applying a topic model specifying 40 topics to the Aljazeera corpus, we identified a generic 'sports' topic and a generic 'culture' topic (including comments to articles about language, science, technology, society) among the other political topics. We selected comments to sports and culture articles by taking the top few hundred comments having the highest probability score for these 'topics', to guarantee that the content was indeed relevant to the domain. Table 6.2 shows the distribution of the final data used for annotation, consisting of 1177 news article comments.

6.1.3.2 Data Characteristics

As mentioned previously, the spans of text used for identifying targeted sentiment are long and complex. The average length of news article comments in the annotated dataset is 51 words, ranging from 1-3 sentences per comment. The data was not

[5] https://www.aljazeera.com/

corrected for spelling errors; we annotated the raw text to avoid any alteration that may affect the interpretation of sentiment. However, it is possible to correct this output automatically, such as with the approach of Farra et al. (2014), or manually.

We performed a manual analysis of 100 article comments from a randomly selected subset of the dataset with the same domain distribution. We found that 43% of the comments contain at least one spelling error including typos, word merges and splits,[6] 15% contain at least one dialect word, 20% contain a run-on sentence not separated by any conjunction or punctuation, and 98% express any sentiment. We believe this is a good dataset for annotation because it is sufficiently challenging, contains real-world data, and includes strong expressions of sentiment covering multiple controversial topics.

6.1.4 Dataset Analysis

This section describes results and analyses of the crowdsourced annotations. We report the inter-annotator agreement at each of the two annotation stages, the distribution of the sentiment of collected targets by domain, and a manual analysis of the resulting target entities. Examples of the final annotations are provided.

6.1.4.1 Inter-annotator Agreement

Task 1: Agreement on Candidate Entities. To compute the agreement between annotators a_1, a_2, and a_3 on identifying important entities in a HIT, we compute the average precision p_{HIT}. p_{HIT} is then averaged over all HITs to obtain the agreement.

$$p_{HIT} = \frac{1}{3}.\left(\frac{\#matches}{\#phrases\_a1} + \frac{\#matches}{\#phrases\_a2} + \frac{\#matches}{\#phrases\_a3}\right) \tag{6.1}$$

[6] We don't count the different variations of *Alef* ا, ي/ى, or ه/ة, forms, which are often normalized during model training and evaluation.

An average precision of 0.38 was obtained using exact matching of entities and 0.75 using subset matching: i.e a match occurs if the three annotators all list a sub-phrase of the same noun phrase. (Recall that the final entities were chosen according to subset agreement.)

The candidate entity agreement numbers are comparable to the target span subset agreement numbers of Somasundaran et al. (2008) in English discourse data, and lower than that of Toprak et al. (2010), who annotated targets in the consumer review domain. We note that besides the language difference, the task itself is different, since it requires annotation of important entities rather than sentiment targets; a lower agreement on this task essentially indicates that fewer entities are being passed on to the next task for consideration as targets, the assumption being that only important entities will be agreed upon by all three annotators. Since we had three rather than two annotators, the agreement using exact match is expected to be low.

Task 2: Sentiment agreement. Table 6.3 shows the annotator agreement for the task of identifying sentiment towards given entities. A majority agreement occurs when 3 out of 5 annotators agree on whether the sentiment towards an entity is positive, negative, or neutral. The agreement (91.8%) is reasonably high. Abdul-Mageed and Diab (2011) have reported overall agreement of 88% for annotating sentence-level Arabic sentiment (as positive, negative, neutral, or objective) using two trained annotators. We note that after assigning our task to only the qualified group of workers, the annotator agreement increased from 80% and 88% in the first two annotation rounds, to 95% in the remaining rounds. Target entities with disagreement over polarity are marked as ambigious, or ‘undetermined’.

Domain	# Entities	Majority Agree (%)
Politics	3853	91.2
Culture	2271	95.8
Sports	1222	87.6
Total	**7346**	**91.8**

Table 6.3: Agreement on entity-level sentiment annotation.

Domain	# Targets	(%) Pos	(%) Neg
Politics	2448	30	70
Culture	1149	48	52
Sports	748	79	21
Total	**4345**	**43**	**57**

Table 6.4: Distribution of sentiment in targets with majority agreement (Pos:Positive, Neg:Negative).

6.1.4.2 Sentiment Distribution

Table 6.4 shows the distribution of the sentiment of non-ambiguous targets by domain. These were sentiment targets targets assigned to positive or negative labels by majority annotator agreement. We can see that the politics and sports domains are biased towards negative and positive sentiment respectively, while targets in the culture domain have a mostly even distribution of sentiment. We also note that overall, 95% of all article comments had at least one target of sentiment, and 41% of these comments had multiple targets with both positive and negative sentiment, indicating the need for fine-grained targeted sentiment analysis of such datasets.

Finally, we found that the majority of targets are composed of 2 words (38% of targets), followed by 1-word targets (25% of targets), 3-word targets (18%), and 4-word targets (9%), while 10% of all targets are composed of more than 4 words.

Observation	Example
Spelling errors **2.5%**	ارادت الشعب *“the people’s will”*
Punctuation **5%**	منتجات ابل. *“Apple’s products.”*
Prep & Conj clitics **8.5%**	لمانشتر يونايتد *“to Manchester United”*
Non-noun phrases **3%**	البرشا بطل الدور الاسباني *“Barcelona (is) the champion of the Spanish league”*
Targets with sentiment **5.5%**	الشعب السوري الحر *“the free Syrian people”*
Propositional entities **3%**	تشجيع الباحثين *“encouraging researchers”*

Table 6.5: Examples of target entity observations.

6.1.4.3 Manual Analysis

We manually examined 200 randomly selected targets from our final dataset, and found a number of observations, many of which are language-specific, that deserve to be highlighted. They are summarized in Table 6.5.

We first note orthographic observations such as spelling errors, which come mostly from the original text, and punctuations attached to targets, which may easily be stripped from the text. The punctuations result from our decision to take the maximal noun phrase provided by annotators.

Prepositional and conjunctional clitics result from Arabic morphology which attaches prepositions such as *l+* ل *(to)* and *b+* ب *(in)*, or conjunctions *w+* و *(and)* to the noun preceding them. They can be separated by tokenization as described in Chapter 5, but we preserve them in the dataset for completeness and apply tokenization during modeling instead.

Non-noun phrases mainly come from nominal sentences specific to Arabic syntax, which lack a linking verb such as ‘is’, making it appear like a noun phrase; these are problematic because they may be interpreted as either noun phrases or full sentences that begin with a nominal. We also observed a number of verbal phrase targets (e.g ‘نبلبل بالديموقراطية’ *‘we confuse democracy’*), but these were very few; the majority of this class of observations comes from verbless nominal phrases.

Targets containing sentiment words appear since sentiment words can be part of the noun phrase. As for propositional entities (e.g *‘I support **encouraging researchers**’*), they result from process nominals which can have a verbal reading (Green and Manning, 2010) but are correctly considered to be nouns. We find that they occur mostly in the culture domain.

We also found from our manual inspection that our final entity spans reasonably corresponded to what would be expected to be targets of sentiment for the topic in context. From our 200 randomly selected targets, we found 6 cases where the polarity towards the noun phrase potentially negated the polarity towards a shorter entity within the noun phrase. However, in most of these cases, the noun phrase resolved from the annotations correctly represents the actual target of sentiment: e.g. ‘*depletion of **ozone***’ ثقب الاوزون (the depletion is the target of discussion, not the ozone), *‘bombing of **houses***’ قصف المنازل, and *‘methodology of teaching **Arabic***’ اسلوب تعليم العربية. We found one case *‘absence of **Messi**’* غياب مسي, labeled negative, where it could be argued that either *Messi* (positive) or his absence (negative) is the correct target. We generally preferred target annotations which correspond to the topic or event being discussed in the context of the comment.

6.1.4.4 Examples

We provide examples of the final annotations, shown in Tables 6.6-6.8. Note that we have preserved all spelling errors in the original Arabic text. As it is common in social media to write long sentences without punctuation, we have added punctuation to the English translation.

Example (1) Domain: Culture	**Article Comment** رغم انتشار **الكتاب الألكتروني** الا ان **الكتاب الورقي** اثبت وجوده. احب **الكتاب المطبوع** .. حتى تقليب صفحاته أجد بها متعة.. والأجمل عند قراءته وهو بين يدي .. لا أحتمل **قراءة الكتاب من خلال الشاشة** .. لا أستطيع الاستمرار في تحمل وهج الضوء والصداع.. الكتاب التقليدي أقراءه في المكتبة في القطار في الطائرة على الشاطئ في الحديقة في اي مكان أرتاح فيه .. لامكان للكتاب الألكتروني في قاموسي.
English Translation	Despite the popularity of **the e-book**, **the paper book** has proven itself. I like **the printed book**... I even find a pleasure in turning its pages ... and it is nice is to read it while it is in my hands ... I cannot stand **reading a book through a screen** ... I cannot bear the glare of light and the headaches...I can read a traditional book in the library on the train in the airplane on the beach in the garden in anywhere I am comfortable .. there is no place for the e-book in my dictionary.
Annotated Targets	**negative:** the e-book الكتاب الالكتروني **positive:** the paper book الكتاب الورقي **positive:** the printed book الكتاب المطبوع **negative:** reading a book through a screen قراءة الكتاب من خلال الشاشة

Table 6.6: Example 1 of target annotations. The original spelling errors are preserved.

Example (1) is from the culture domain. We see that it summarizes the writer's opinions towards all important topics regarding 'e-books' and 'paper books'. Ideally, the annotators should also have marked *traditional book* الكتاب التقليدي as a positive target.

Example (2) lists an entity that doesn't appear in the text *'(to) the Arab team the world cup'* للمتخب العربي المونديال; this likely results from an error in Task 1 where the phrase got picked up as the maximal common noun phrase. The annotator might have meant that 'Arab team in the world cup' is a topic that the writer feels positively about; however, our current annotation scheme only considers entities that strictly appear in the text. We also see that annotators disagreed on the polarity of the propositional entity *'either team qualifying'* تأهل الفريقين, likely because they were not sure whether it should be marked positive. In addition, this example contains an over-generated target *'world cup'* المونديال, which would have been best marked as neutral.

Example (3) is from the politics domain. It correctly annotates multiple references of the Iraqi government' and captures the sentiment towards important entities in the text. The target *'the only neighboring country'* الدولة الجارة الوحيدة can be considered

Example (2) Domain: Sports	**Article Comment** **المنتخبان المصري والجزائري** هما منتخبان قويان. والدعم الدي حضي به **المنتخب الجزائري** بالمناسبة جعل الكل مثوثر ولايوجد فرق في تأهل الفريقين و ا تمنى ان يتأهل **الفريق الجزائري** الي **المنونديال** لأنّني احب **الفريق الجزائري** الى جانب المنتخب المصري . والمهم التمثيل الجيد و ا تمنى ان يكون **للمتخب العربي** احسن تمثيل في **المنونديال** .
English Translation	**The Egyptian and Algerian teams** are strong teams. The support gained by the **Algerian team** for this occasion has made everyone nervous and there is no difference in **either team qualifying** and I hope that **the Algerian team** gets qualified to **the world cup** because I like **the Algerian team** alongside the Egyptian team. The important thing is good representation and I hope that **the Arab team** will be best represented in **the world cup**.
Annotated Targets	**positive:** The Egyptian and Algerian teams المنتخبان المصري والجزائري **positive:** the Algerian team المنتخب الجزائري **positive:** the Algerian team الفريق الجزائري **positive:** the world cup المنونديال **positive:** (to) the Arab team the world cup للمتخب العربي المنونديال **undetermined:** either team qualifying تأهل الفريقين

Table 6.7: Example 2 of target annotations. The original spelling errors are preserved.

an over-generation; a better interpretation might be to consider this phrase part of the sentiment expression itself. Nonetheless, this extra annotation may provide helpful information for future modeling.

Example (3) Domain: Politics	**Article Comment** مع الاسف **الحكومة العراقية** لا يفتهم من السياسة شيء لأن **الدولة الجارة الوحيدة** التي تربطنا معها اكثر من مصالح من الموارد الطبيعية كالمياه الى مصالح صناعية هي **تركيا** فعلينا ان نقوي علاقتنا معها لانها اصبحت تنافس الدول الاوربية لنستفاد منها ولكن **حكومة المالكي الفاشلة** لا يهمهم التطور وقد رجع **العراق** بظل هؤلاء مئات السنين الى الخلف.
English Translation	Unfortunately **the Iraqi government** understands nothing of politics because **the only neighboring country** with whom we have ties that are not just based on interests - such as natural resources like water and industrial interests - is **Turkey**, so we have to strengthen our relationship with it because it is now a competitor with European nations, we should benefit from it but **Maliki's failed government** cares nothing for progress and **Iraq** has gone back hundreds of years because of these people.
Annotated Targets	**negative:** the Iraqi government الحكومة العراقية **positive:** the only neighboring country الدولة الجارة الوحيدة **positive:** Turkey تركيا **negative:** Maliki's failed government حكومة المالكي الفاشلة **negative:** Iraq العراق

Table 6.8: Example 3 of target annotations. The original spelling errors are preserved.

We generally found that the annotations correctly covered sentiment towards essential targets and mostly complied with our definition of entities. The annotations contain some errors, but these are expected in a crowdsourcing task, especially one that relies to a degree on some subjective interpretation. We noticed that annotators tended to over-generate targets rather than miss out on essential targets. The annotation of these secondary targets may prove useful for future modeling tasks.

6.2 Open-Domain Target and Sentiment Models

This section describes the targeted sentiment models developed for identifying entity targets along with their sentiment in the open-domain dataset presented in Section 6.1. As described previously, the open-domain task consists of identifying all targets towards which sentiment (positive or negative) is expressed in the short document, along with the polarity associated with each target. Targets of sentiment can include any nominal, and are not restricted to named entities, but they must be explicitly mentioned in the text in order to be selected by the system.

We develop two sequence labeling models, a target-specific model and a sentiment-specific model. The models try to learn syntactic relations between candidate entities and sentiment words, but they also make use of (1) Arabic morphology and (2) entity semantics. The use of morphology allows the model to capture all 'words' that play a role in identification of the target, while the use of entity semantics allows the model to group together similar entities which may all be targets of the same sentiment; for example, if comments express negative sentiment towards the 'United States', they *may* also express negative sentiment towards 'America' or 'the American president' - this hypothesis is to be tested by the use of entity clusters.

Our results show that here as well, morphology matters when identifying entity targets and the sentiment expressed towards them. We find for instance that the attaching Arabic definite article *Al+* ال is an important indicator of the presence of a target entity and splitting it off boosts recall of targets, while sentiment models perform better when less tokens are split.

Sections 6.2.1 and 6.2.2 describe the targeted sentiment models and linguistic decisions made for Arabic. Sections 6.2.3 and 6.2.4 present experiments and results, where models are evaluated under a variety of resource settings, including a high-resource setting where rich linguistic features are available, and a more low-resource setting where only monolingual word vectors are available and are used for building cluster features. A detailed analysis of errors, shown in Section 6.2.5, reveals that the task generally entails hard problems.

6.2.1 Sequence Labeling Models

We chose to model the data using Conditional Random Fields (CRF) (Lafferty et al., 2001), for their ability to allow engineering of rich linguistic features and their past success in tasks such as entity identification and sequence labeling. Moreover, CRF models are global, in that when a decision is being made about the tag for a given timestep of the sequence, the entire candidate sequence is scored, as opposed to the state at a given timestep. This property is helpful for identifying global patterns in the sequence, such as locations of targets or sentiment. (While perhaps not as powerful as a deep learning model, this property gives the CRF an advantage over the standard bidirectional LSTM, although there are other models, such as biLSTM with attention to different hidden states (Bahdanau et al., 2014), and biLSTM-CRF (Huang et al., 2015), which may certainly be explored for this task in future studies.) Two linear chain CRF models were constructed:

1. **Target Model.** This model predicts a sequence of labels $\mathbf{E} = e_1, e_2, ...e_n$ for a sequence of input tokens $\mathbf{x} = x_1, x_2, ...x_n$, where:

$$e_i \in \{T(target), O(not\_target)\}$$

and each token x_i is represented by a feature vector f_{it}. A token is tagged T if it is part of a target; a target can contain one or more consecutive tokens.

2. **Sentiment Model.** This model predicts a sequence of labels $\mathbf{S} = s_1, s_2, ...s_n$ for the sequence $\mathbf{x}$, where:

$$s_i \in \{P(positive), N(negative), \emptyset(neutral)\}$$

and each token x_i is represented by a feature vector $f_{is}; e_i$, where $e_i \in \{T, O\}$ is pipelined from the output of the target model. Additionally, this model has the constraint:

$$if E_i = T, S_i \in \{P, N\} \tag{6.2}$$

and otherwise,

$$S_i = \emptyset \tag{6.3}$$

The last constraint ensures that only targets of sentiment can be tagged positive or negative, and non-targets are always assigned a neutral label. The target and sentiment models are trained independently. Thus, if target keywords are already available for the data, the sentiment model can be run without training or running the target model. Otherwise, the sentiment model can be run on the output of the target predictor. The sentiment model uses knowledge of whether a word is a target and utilizes context from neighboring words in the sequence in order to predict sentiment polarities towards the targets. An example sequence is shown in Table 6.9, where the target 'the dictator' is an entity towards which the writer implicitly expresses negative sentiment.

The	**dictator**	**is**	**destroying**	**his**	**country**
T	T	O	O	O	O
N	N	∅	∅	∅	∅

Table 6.9: Example of CRF training data.

6.2.2 Arabic Morphology and Linguistic Features

This section describes the linguistic features and language specific considerations that are applied for creating the data used to train the targeted sentiment models. The features include different morphological segmentation techniques and representations, sentiment lexicon features, syntactic dependencies, base phrase chunks and named entities. In addition to rich linguistic feature represenations, we also consider a more low-resource scenario where only word vector clusters are available.

6.2.2.1 Arabic Morphology

As described in Chapter 5, Arabic exhibits complex concatenative and inflectional morphology. Concatenative morphology is exhibited through the attachment of clitics and affixes to the beginning and end of the word stem, making words complex. For example, in the sentence فاستقبلوها *fAstqblwhA*, 'So they welcomed her', the discourse conjunction 'so' +ف, the sentiment target 'her' ها+, sentiment source 'they' وا+ , and the expression of sentiment itself ('welcomed' استقبل) are all collapsed in the same word. (Note that we do not consider the identification of pronominal targets through coreference; some of our early experiments attempted this, but achieved no gains as the task required a more complex co-reference system than what was available for Arabic at the time of development.)

Clitics, such as conjunctions +و *w+*, prepositions +ب *b+*, the definite article ال+ *Al+* (all of which attach at the beginning), possessive pronouns and object pronouns ه+ and ها+ 'his/her' or 'him/her' (which attach at the end) can all function as individual words. Thus, they can be represented as separate tokens in the CRF model.

Similar to Chapter 5, we use the morphological analyzer MADAMIRA (Pasha et al., 2014) tokenize words using multiple schemes. We consider the following two schemes:

- **D3**: the three-level declitization scheme which splits off conjunction clitics, particles and prepositions, *Al+*, and all the clitics which attatch at the end.
- **ATB**: the Penn Arabic Treebank tokenization scheme, which separates all clitics above except the definite article *Al+* ('the'), which it keeps attached.

In addition to segmentation schemes, we address inflectional morphology by incorporating detailed part-of-speech (POS) features. Each token is assigned a POS produced by the morphological analyzer; for clitic tokens, we also assign POS tags such as 'determiner' for *Al+* or 'third person masculine posesssive pronoun' for ه+ 'his'.

To represent word forms, we consider both the sparse surface word and the lemma. Figure 6.5 shows the different possible representations of words and clitics used in the CRF model, using the example 'with help from the government'. All lexical and POS features are added to both our target model and sentiment model.

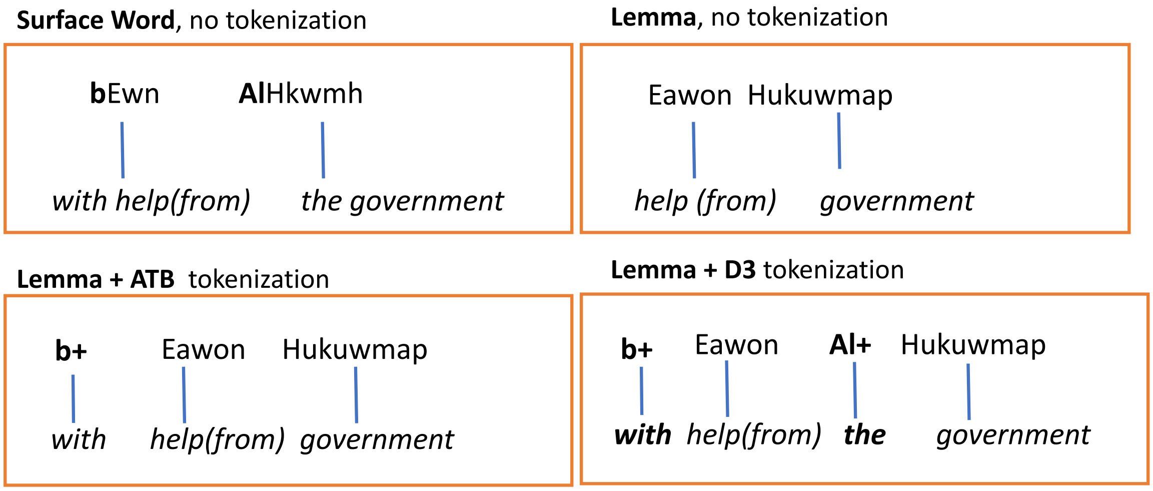


Figure 6.5: Tokenization and word representation schemes used in target and sentiment models.

6.2.2.2 Sentiment Features

We consider three sentiment lexicons for creating sentiment features:

1. *SIFAAT*, a manually constructed Arabic lexicon of 3982 adjectives (Abdul-Mageed and Diab, 2011).
2. *ArSenL*, an Arabic lexicon developed by linking English Sentiwordnet with Arabic Wordnet (Black et al., 2006) and an Arabic lexical database (Badaro et al., 2014).
3. The English MPQA lexicon (Wilson et al., 2005), where we look up words by matching on the English glosses produced by the morphological analyzer MADAMIRA.

We evalute the three lexicons separately and use the best performing lexicon in the CRF model, while all lexicons are used when creating lexical baselines. For the target model, we create token-level binary features representing subjectivity (presence or absence of any positive or negative sentiment), and for the sentiment model, we create both subjectivity and polarity features.

We also create a feature specifying respectively the subjectivity or polarity of the parent word of the token in the dependency tree in the target or sentiment model, whereby syntactic dependencies described in the following section.

6.2.2.3 Syntactic Dependencies

We ran the Columbia Arabic Treebank (CATiB) dependency parser (Shahrour et al., 2015) on our data in order to create features that identify syntactic relations between potential target entities and neighboring words in the sequence. CATiB uses a number of intuitive labels specifying the token's syntactic role - such as subject 'sbj', object 'obj', modifier 'mod', or 'idf' for the Arabic 'idafa' construct which indicates posessiveness (e.g رئيس الحكومة 'president of government') - as well as its part of speech role.

In addition to dependency features specifying the sentiment of parent words, we create dependency features specifying the syntactic role of the token in relation to its parent, and the path from the token to the parent, e.g *nom_obj_verb* (nominal that is an object of a verb) or *nom_idf_nom* (nominal related to parent nominal through idafa), as well as the sentiment path from the token to the parent, e.g *nom*(neutral)*_obj_vrb*(negative) (neutral nominal that is an object of a verb with negative sentiment).

6.2.2.4 Chunking and Named Entities

We create features specifying base phrase chunks (BPC) - these are simple sentence chunks indicating spans of noun, verb, and prepositional phrases - and named entity tags (NER) for each token. We use these features based on the hypothesis that they will help define the spans for entity targets, whether they are named entities or generic noun phrases. Both BPC and NER are produced by MADAMIRA.

We refer to the sentiment and target models that utilize Arabic morphology, sentiment, syntactic relations and entity chunks as BEST-LINGUISTIC.

6.2.2.5 Word Embedding Clusters and Entity Semantics

While most of the previously described features are applicable in a high-resource scenario where syntactic and morphological tools are available, it is also possible to specify discrete features that can be more easily made available whether or not such tools are available for the language. We consider word cluster features, based on the hypothesis that similar entities which occur in the context of the same topic or the same larger entity are likely to occur as targets alongside each other and to have similar sentiment expressed towards them. Such entities may repeat frequently in an article comment even if they do not explicitly or lexically refer to the same person or object. For example, someone writing about American foreign policy may frequently refer to entities such as ‘the United States’ , ‘America’, ‘the Americans’, or ‘the West’. Such entities can cluster together semantically and it is possible that a comment expressing positive or negative sentiment towards one of these entities may also express the same sentiment towards the other entities in this set.

Moreover, cluster features serve as a denser feature representation compared to the sparser full vocabulary and they have been used effectively for named entity tagging tasks, e.g by Zirikly and Hagiwara (2015). Such features can benefit the CRF where a limited amount of training data is available for target entities.

To utilize the semantics of word clusters, we build monolingual word embedding vectors using the skip-gram method (Mikolov et al., 2013) and apply K-Means clustering (MacQueen, 1967), with Euclidean distance as a metric. Euclidean distance serves as a semantic similarity metric and has been commonly used as a distance-based measure for clustering word vectors. While Brown clusters were used in Chapter 4 and may also be used for targeted sentiment models, we chose K-Means clusters in order to mimic the named entity tagging experiments of Zirikly and Hagiwara (2015). We varied the number of clusters and used the clusters as binary features in our target and sentiment models.

6.2.3 Experiments

We ran experiments evaluating the target and sentiment identification models individually, as well as the full pipeline that predicts both target entities and the sentiment towards them. Our experiments assess the following:

1. The effect of different morphological schemes and word representation forms, on identifying targets and their sentiment in Arabic.

2. The effect of high-resource, rich linguistic features, on the performance of the target and sentiment models.
3. The effect of low-resource word embedding clusters on the performance of the target and sentiment models.
4. The overall performance of the full pipeline on the open-domain task applied to short Arabic documents.

6.2.3.1 Setup and Configurations

We used CRF++ (Kudo, 2005) to build linear-chain sequences for the target and sentiment identification models. We used a context window of +/-2 neighboring words for all features except the syntactic dependencies, where we used a window of +/-4 to better capture syntactic relations in the posts. For the sentiment model, we additionally included the context of the previously predicted label, to avoid predicting consecutive tokens with opposite polarity.

The Sentiwordnet-based lexicon ArSenL uses real-valued scores for representing the sentiment of words; we discretized it by using a threshold of *t=0.2*.

The vectors used for creating word embedding clusters were built on Arabic Wikipedia[7] on a corpus of 137M words resulting in a vocabulary of 254K words. We used word2vec[8] for building and clustering 200-dimensional word vectors. We preprocessed the corpus by tokenizing (using the schemes described in section 6.2.2) and lemmatizing before building the word vectors. We varied the number of clusters k between 10 (25K words/cluster) and 20K (12 words/cluster).

[7] https://dumps.wikimedia.org/arwiki/20160920/arwiki-20160920-pages-articles.xml.bz2

[8] https://github.com/dav/word2vec

6.2.3.2 Data

We use the dataset we created, described in Section 6.2, and divided it into a training set (80%), development set (10%), and blind test set (10%), all of which are representative of the three domains: politics, sports, and culture. The data contains ambiguous or 'undetermined' targets where a majority of annotators assigned positive or negative sentiment, but did not agree on the polarity. We used these targets for training our target identification model, but discarded them when training our sentiment identification model.

We cleaned and preprocessed the data and the targets, making sure that all target entities appear explicitly in the text and discarding those which do not. There are 4886 targets distributed as follows: 38.2% positive, 50.5% negative, and 11.3% ambiguous. We make this cleaned data as well as the splits publicly available in addition to the original dataset.

Since our models do not require parameter tuning, we evaluated all our experiments on the split reserved for the development set, which contains 116 posts and 442 targets, and we retain the held-out test set for future experiments.

6.2.3.3 Baselines

We incorporate the following baselines:

1. **All-NP**: For evaluating the identification of targets, we follow work in English (Deng and Wiebe, 2015a) and use the 'all-NP' baseline, where all nouns and noun phrases in the post are predicted as important targets. The Stanford Parser (De Marneffe and Manning, 2008) is used to identify noun phrases.
2. **Majority**: For evaluating the identification of sentiment towards targets, we use the majority baseline, which always predicts negative.
3. **Lexicon**: For evaluating the identification of sentiment towards targets, we

also use a lexicon baseline evaluated in the case of the three lexicons: manually created (SIFAAT), Sentiwordnet-projected (ArSenL), and English-translated (MPQA). The strong lexicon baseline splits the article comment into sentences or phrases by punctuation, finds the phrase that contains the predicted target, and returns positive if there are more positive words than negative words, and negative otherwise. These baselines are similar to the methods of previously published work for Arabic targeted sentiment (Al-Smadi et al., 2015; Obaidat et al., 2015; Abu-Jbara et al., 2013).

4. **Topical**: This baseline addresses the uneven distribution of sentiment in our dataset based on topics (bias towards negative in politics topics and positive in sports topics), which could potentially be a confounding factor for the model. The topical baseline assigns negative sentiment to all targets retrieved from the politics domain, positive sentiment to all targets retrieved from the sports domain, and defaults to the prediction of the MPQA Lexicon baseline otherwise.

6.2.3.4 Metrics

1. For evaluating the identification of targets, we use Target F-measure, which is determined by computing the recall of and precision of predicted targets by matching with the gold annotated targets. We match targets based on 'subset' (similar to matching schemes used by Yang and Cardie (2013), Irsoy and Cardie (2014)); if either the predicted or gold target tokens are a subset of the other, the match is counted when computing F-measure. Overlapping matches that are not subsets do not count (e.g 'Egypt's position' and 'Syria's position' do not match). In the case of multiple mentions of the same entity in the post, any mention will be considered correct.
2. For evaluating the identification of sentiment towards targets, we compute accuracy 'Acc', as well as the positive class F-measure 'F-pos' and the negative

class F-measure ‘F-neg’. These are only evaluated on *correctly predicted* targets. Since the target and sentiment models are trained separately, these scores are meant to reflect how the targeted sentiment model would perform if targets were already known.

3. For evaluating the end-to-end task of target and sentiment identification, we use ‘F-all’, the overall F-measure comparing correctly predicted targets with correct sentiment compared to the total number of polar targets.

6.2.4 Results

Table 6.10 shows baseline results using the ‘all-NP’ target baseline and the five sentiment identification baselines: the majority baseline, three sentiment lexicon baselines, and topical baseline. Table 6.11 shows the performance of the CRF models using morphological representation schemes: surface word representation (no token splits), lemma represenation, lemma with ATB clitics (includes all tokens except *Al+*), and lemma with D3 clitics (includes all token splits). (See Figure 6.5 for a representation of these schemes.) Table 6.12 shows the performance of the high-resource best-linguistic model under various tokenization scenarios combining the D3 and ATB schemes.

Significance thresholds are calculated for the best performing systems using the approximate randomization test Yeh (2000) for target recall, precision, F-measure, accuracy and overall F-Measure. Significance over the method in the previous row is indicated by *(p <0.05), **(p <0.005), **(p <0.0005). A confidence interval of almost four F-measure points is required to obtain p <0.05.

6.2.4.1 Baseline Performance

From Table 6.10, we see that as expected, the ‘All-NP’ baseline has near perfect recall (98.4) and low precision (29.2) in predicting important targets, since it assumes that

	Target			Sentiment			
	Recall	Precision	F-score	F-pos	F-neg	Acc	F-all
Majority	98.4	29.2	45.0	0.0	72.4	56.8	12.4
ArSenL	98.4	29.2	45.0	50.6	64.3	58.6	12.7
SIFAAT	98.4	29.2	45.0	61.0	58.0	59.5	13.1
MPQA	98.4	29.2	45.0	**67.0**	63.7	65.4	14.2
Topical	98.4	29.2	45.0	56.4	**76.8**	**68.7**	**14.9**

Table 6.10: Target and sentiment identification results using baselines. The 'all-NP' baseline is applied for identifying targets for all five sentiment baselines.

	Target			Sentiment			
All-NP	Recall	Precision	F-score	F-pos	F-neg	Acc	F-all
Surface + POS	41.0	**60.6**	48.9	62.2	73.6	68.9	32.6
Lemma + POS	48.2**	60.5	53.7*	**65.4**	**77.6**	**72.8**	38.1**
+**ATB** tokens	52.4*	59.5	55.7	61.3	75.7	70.1	**38.2**
+**D3** tokens	**59.6****	55.7*	**57.6**	64.1	73	69.2	36.1

Table 6.11: Target and sentiment identification results using different morphological representations: surface word, lemma, lemma+ATB tokenization, and lemma+D3 tokenization. Significance over the method in the previous row is indicated.

	Target			Sentiment			
	Recall	Precision	F-score	F-pos	F-neg	Acc	F-all
ATB	53.0	**62.1**	57.2	68.6	79.4	75.1	40.7
D3	64.2***	58.8	61.4*	62.7	75.6	70.5*	39.1
D3+ATB	63.7	58.8	61.4	67.7	**80.0**	75.4***	43.1***
+clusters	**66.2**	57.8	**61.8**	**70.0**	**80.0**	**76.0**	**44.2**

Table 6.12: Performance of best-linguistic model with different Tokenization Schemes: ATB, D3, D3+ATB, and word embedding clusters. Significance over the method in the previous row is indicated.

every noun phrase is a target of sentiment.

For predicting sentiment towards correctly predicted targets, we observe that the gloss-translated MPQA lexicon outperforms the majority baseline and the two other Arabic lexicons, with a targeted sentiment accuracy of 65.5, positive class F-measure of 67.0 and negative class F-measure of 63.7. The hit rate of MPQA, which is composed of a combination of manually labeled and automatically generated clues, is higher than that of the smaller, manually-labeled SIFAAT, and it is more precise than the automatically generated Sentiwordnet-based lexicon ArSenL. The performance of MPQA is, however, reliant on the availability of high-quality English glosses. Early experiments revealed that MPQA features consistently perform upon integration in CRF models, so we use the gloss-translated lexicon to create features for the BEST-LINGUISTIC model.

The topical baseline outperforms the MPQA lexical baseline with a targeted sentiment accuracy of 68.7, reflecting the label bias of topics in the dataset. However, its performance is very different than MPQA when considering positive vs. negative targets, whereby the MPQA baseline does better with positive targets (67.0 vs. 56.4 positive F-measure) and the topical baseline does better with negative targets (76.8 vs. 63.7 negative F-measure). This indicates that the MPQA lexicon predicts sentiment independently of topics.

The overall best F-measure performance using the All-NP baseline is 14.9, which provides a measure of the difficulty of the end-to-end task.

6.2.4.2 Evaluating Morphological Representations

From Table 6.11, we see that using the lemma representation easily outperforms the sparser surface word representation (increase in target F-measure from 48.9 to 53.7, in sentiment accuracy from 68.9 to 72.8, and in overall F-measure from 32.6 to 38.1).

The addition of tokenized clitics further improves target identification upon morphological representations which only use the word form, leading to a target F-measure of 55.7 and an overall F-measure of 38.1 using ATB tokens. Moreover, upon using the D3 decliticization method, we observe a significant increase in recall of sentiment targets over the ATB representation, leading to a target recall of 59.6 and a target F-measure of 57.6. This interesting result shows that the presence of the Arabic definite article 'Al+' is an important indicator of the presence of a target entity; thus, even if an entity is not named, the definite article indicates that it is *known* entity and is likely more salient or important to the topic of the text.

The more tokens are split off, the more targets are recalled, although this comes at the cost of a decrease in sentiment performance, where the lemma representation has the highest sentiment accuracy (72.8) and the D3 representation has the lowest (69.2) after surface word (68.9). It is possible that the addition of extra tokens in the sequence (which are function words and have not much bearing on semantics) generates noise with respect to the sentiment model.

All models significantly improve the baselines on F-measure; on sentiment accuracy, the surface word CRF does not significantly outperform the MPQA baseline. Similarly, the surface word CRF only narrowly outperforms the topical baseline on identifying sentiment. However, the addition of morphological representations allows the CRF models to go beyond topical predictions, with the lemma model doing substantially better than the topical baseline on predicting positive targets (65.4 vs. 56.4) and nearly one point in F-measure better on predicting negative targets (77.6 vs. 76.8).

6.2.4.3 Evaluating the Best Linguistic Model

Table 6.12 shows the performance of the BEST-LINGUISTIC model, which in addition to lemma and part of speech features, also uses named entities, base phrase chunks, syntactic dependencies, and sentiment lexicon features. The rich linguistic model was run using both ATB and D3 tokenization schemes, and then using a combined ATB+D3 scheme where D3 tokens were used for predicting targets and the extra clitics were removed before piping in the output to the sentiment model. This combined scheme results in the best results overall: F-measure of 61.4 for identifying targets, accuracy of 75.4 for identifying sentiment and overall F-measure of 43.1.

Adding the richer linguistic resources results in both improved target precision, recall, and sentiment scores compared to using only lemma and POS features, with F-measure for positive targets reaching 67.7 for positive targets and 80.0 for negative targets. Comparison with the topical baseline shows clearly that this model uses a lot more than topical information when predicting sentiment towards targets.

The last row shows the best linguistic model D3+ATB upon integrating cluster features (the best result was obtained for k=8000, or about 30 words per cluster). Adding the clusters results in a small improvement to the target and F-measure scores for the best linguistic model, with a target F-measure of 61.8 and overall F-measure of 44.2. We observe that it becomes more difficult to improve on the rich linguistic model using word clusters, which are more beneficial for low resource scenarios, as will be shown in the following section.

Our results are comparable to published work for most similar tasks in English: e.g Yang and Cardie (2013) who reported target subset F-measure of ˜65, Pontiki et al. (2014) where best performing SemEval systems reported 70-80% for sentiment given defined aspects, and Mitchell et al. (2013); Deng and Wiebe (2015a) for overall F-measure; we note that our tasks differ as described in Chapter 2.

6.2.4.4 Evaluating Word Embedding Clusters

Figures 6.6-6.9 show the performance of different morphological representations when varying the number of word vector clusters k. (Higher k means more clusters and fewer entities per semantic cluster.) In this lower-resource setting, only the word form, POS, and cluster features are integrated in the sequence models.

Adding cluster features tends to further boost the recall of important targets for all morphological schemes, while more or less maintaining precision. The difference in different schemes is consistent with the results of Table 6.11; the D3 representation maintains the highest recall of targets, while the opposite is true for identifying sentiment towards the targets. The ATB representation shows the best overall F-measure,

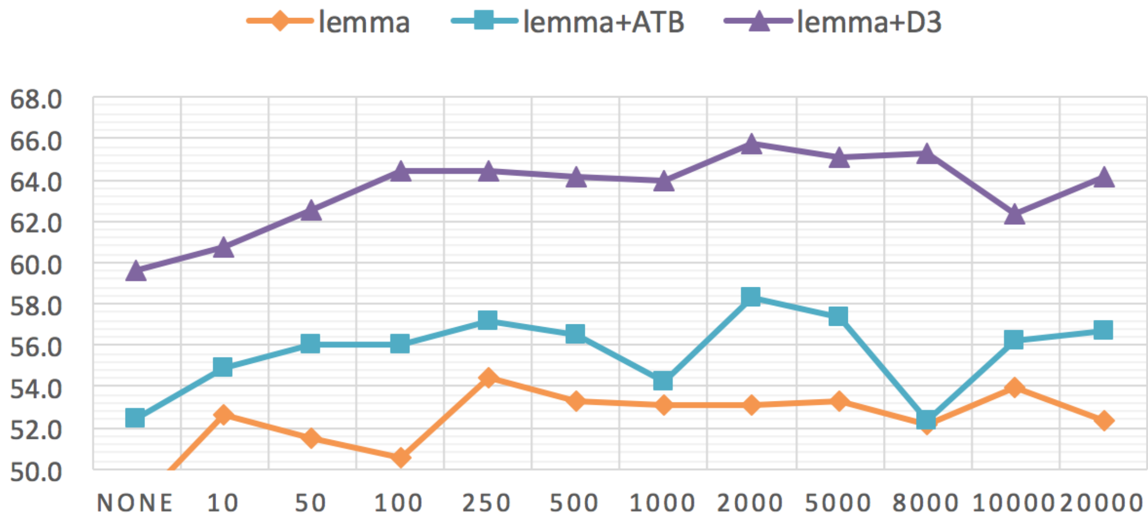


Figure 6.6: Target recall vs clusters.

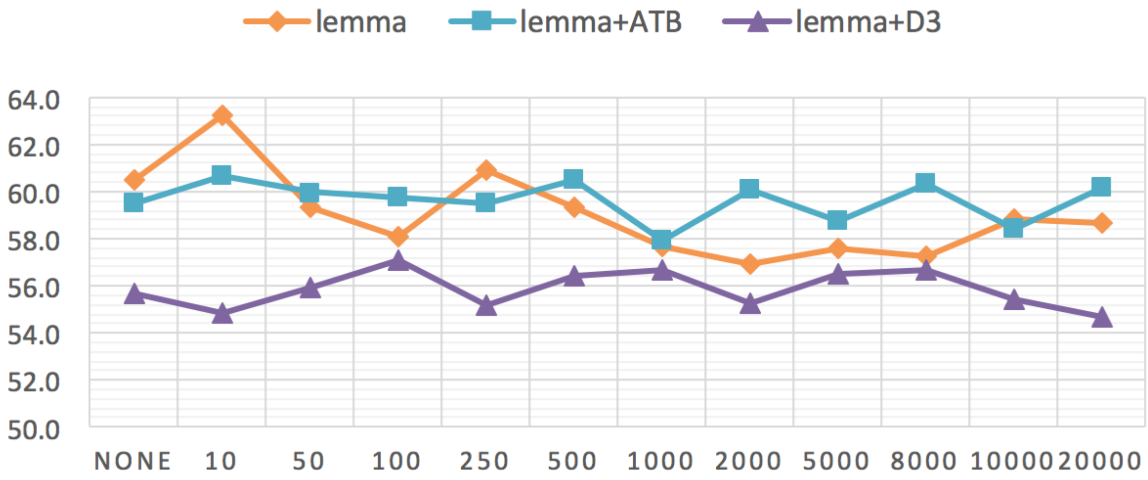


Figure 6.7: Target precision vs clusters.

peaking at 41.5 using k=250 (compare with 38.2 using no clusters); however, it recalls much fewer targets than the D3 representation.

The effect of clusters on predicting sentiment is less clear; it seems to benefit the D3 and ATB schemes more than lemma. The effect of clusters on predicting sentiment is not far in line from our cross-lingual results in Chapter 4; where word clusters resulted in some improvements to sentiment identification but where those improvements were not especially consistent or notable.

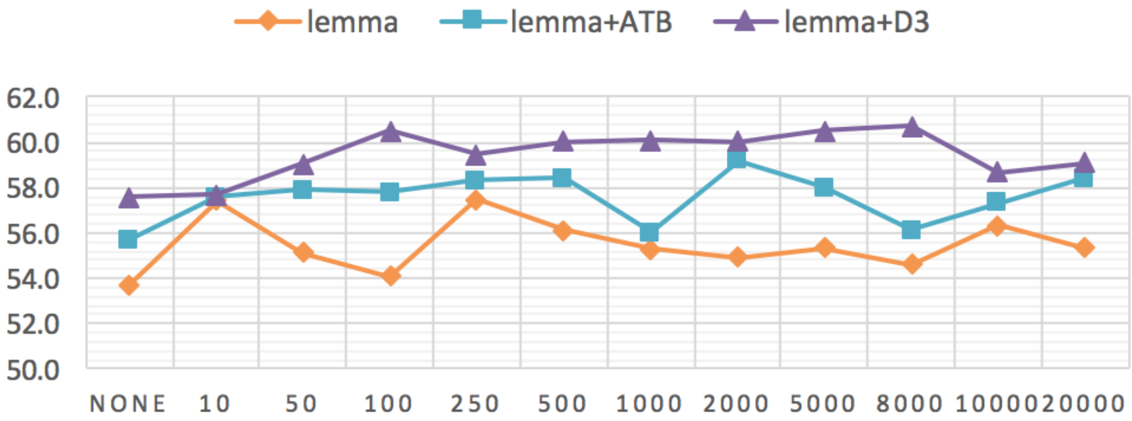


Figure 6.8: Target F-score vs clusters.

The best value of k is k=10 when using lemma, k=250 for lemma+ATB, k=500 for lemma+D3, with corresponding F-all values of 40.7, 41.5, and 39.1 respectively. Performance does not reach that of the best linguistic model, but it still achieves significant boosts in recall and F-measure, bringing it closer to the rich linguistic model. In general, the cluster performances tend to peak at a certain value of k which balances the reduced sparsity of the model (fewer clusters) with the semantic closeness of entities within a cluster (more clusters).

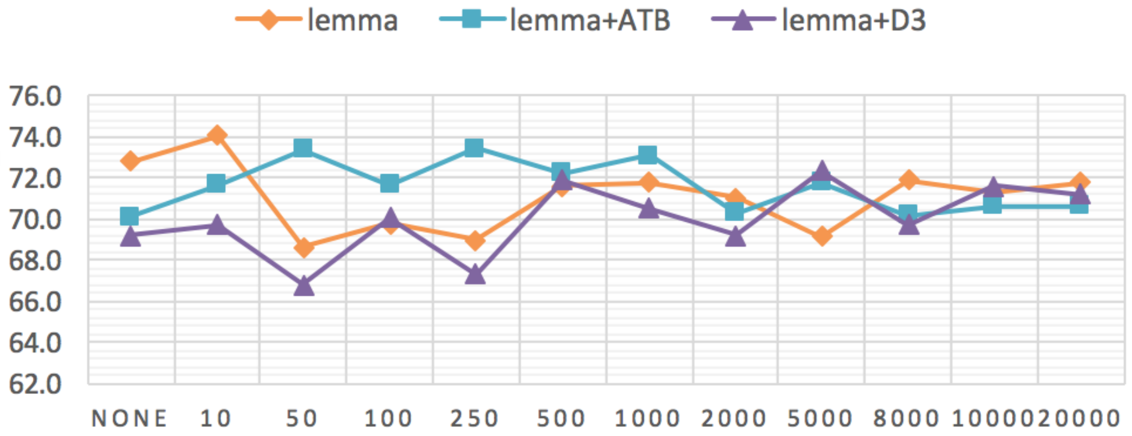


Figure 6.9: Sentiment accuracy vs clusters.

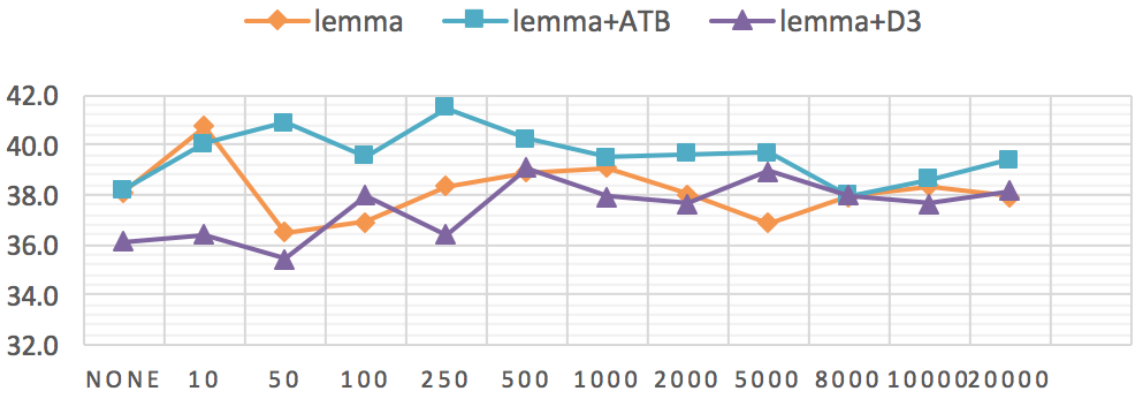


Figure 6.10: Overall F-score vs clusters.

6.2.5 Error Analysis

We analyzed the output of the rich linguistic target and sentiment identification models, and observed a number of kinds of errors, listed below and shown in Table 6.13. Some errors, such as implicit sentiment (requiring inference) and annotation error, overlap with errors observed in Chapter 5 when analyzing our cross-lingual model.

1. **Implicit Sentiment**: This was the most common kind of error observed. Article comments frequently expressed complex subjective language without using sentiment words, often resorting to sarcasm, metaphor, and argumentative language. We also observed persistent errors where positive sentiment was identified towards an entity because of misleading polar words; e.g *minds* العقول was consistently predicted to be positive even though the comment in question was using implicit language to express sarcasm and negative sentiment; the English gloss is *brains*, which appears as a positive subjective word in the MPQA lexicon. The comments also contained cases of complex coreference where subjective statements were at long distances from the targets they discussed.
2. **Annotation Errors:** Our models often correctly predicted targets with reasonable sentiment polarities which were not marked as important targets by

Error Type	Example
Implicit Sentiment	What has [**Messi**]**-** done to take it ???? World cup and not a single goal only the Spanish cup and KFC ads and Pepsi ads ??? so whoever runs more ads is the one who wins it... (*predicts Messi as a positive target*)
Annotation Errors	I hope [**the great and trailblazing Aljazeera**]+ will open this door. And thank you to the great Mr. Amaira and to the great Aljazeera. *(predicts Mr. Amaira as a target)*
Sentiment Lexicon Misses	[**the Syrian revolution**]+ continues by the will of God (*predicts revolution as negative*)
Secondary Targets	[**Jealousy exists**]**-** between [**the Arab regimes**]**-** ...and I am sure that [**Egypt+**] will rise with the help of God.... *(misses ‘Egypt’ and ‘money from the gulf’)*

Table 6.13: Errors, with shortened excerpts and translated examples, made by best-linguistic target and sentiment identification models. Gold annotated targets are shown in brackets [] with ’+’ (positive) or ‘-’ (negative).

annotators; this points to the subjective nature of the task.

3. **Sentiment Lexicon Misses:** These errors resulted from mis-match between the sentiment of the English gloss and the intended Arabic meaning, leading to polar sentiment being missed. For example, in the provided excerpt, while ‘the revolution’ is correctly identified as a target, the original Arabic word is مستمرة ‘persistent’, which has a positive meaning, but it gets translated to ‘continuous’, which has a neutral sentiment in the MPQA lexicon, and therefore the sentiment towards the target is incorrectly predicted. (This particular example contains other targets, such as ‘regime’ which were correctly identified as negative by the model.)

Example 1
Till when will [**the world**]**-** wait before it intervenes against these [**crimes against humanity**]**-** committed by this [**criminal bloody regime**]**-** which will not stop doing that... because its presence has always been associated with oppression and murder and crime... But now it's time for it to disappear and descend into [**the trash of history**]**-**.
Output **the world:**neg **crimes:**neg **criminal bloody regime:**neg **the trash of history**:neg
Example 2
[**Malaysia**]+ is considered the most successful country in Eastern Asia, and its economic success has spread to other [**aspects of life in Malaysia**]+, for its [**services to its citizens**]+ have improved, and there has been an increase in [**the quality of its health and educational and social and financial and touristic services**]+, which has made it excellent for foreign investments.
Output **Malaysia:**pos **health:**pos **educational and social:**neg **financial**:neg

Table 6.14: Good and bad examples of BEST-LINGUISTIC target and sentiment identification output (pos: positive, neg:negative). Gold annotations for targets are provided in the text with '-' or '+' reflecting negative and positive sentiment towards targets.

4. **Secondary Targets:** The data contains multiple entity targets and not all are of equal importance; the majority of targets missed by the model are secondary targets. Out of the first 50 posts manually analyzed on the evaluation set, we found that in 38 out of 50 cases (76%) the correct *primary* targets were identified (the most important topical sentiment target(s) addressed by the post); in 4 cases, a target was predicted where the annotations contained no polar targets at all, and in the remaining cases the primary target was missed. Correct sentiment polarity was predicted for 31 out of the 38 primary correct targets (81.6%). In the last example (which is an excerpt from Figure 6.1), the targeted sentiment models identify all targets and sentiment correctly except 'Egypt' and 'money from the gulf', which it misses.

In general, our analysis showed that the system does well on article comments where targets and language expressing sentiment are well formed, but that the important target identification task is difficult and made more complex by the long and repetitive nature of the text. Table 6.14 shows two examples of the output of the best linguistic models, the first on more well-formed text, where the models correctly predict all targets and sentiment, and the second on text that is more difficult to parse, where the models make a number of errors.

6.3 Situation Frame Task

In the open-domain targeted task for Arabic, the goal was to identify entity targets and sentiment in documents, but the targets had to be explicitly mentioned in the text. In the situation frame task, a 'target' is an abstract concept that is not explicitly mentioned in the text, but that can be inferred from the content.

The goal of the situation frame task is to identify *whether* sentiment is expressed, as well as the *polarity* of the sentiment expressed, towards situations, which can include situations of need, such as 'shelter', or 'medicine', or issue-based situations, such as 'crime' or 'regime change'. While the situation frame task bears similarities to the problem of detecting sentiment towards aspects (e.g good or bad 'service' in a restaurant review), or stance towards an issue (e.g for or against 'gun control'), the situation frame problem is different in that it presents several new challenges:

- A situation is identified not only by its topic or *frame type* (e.g 'medicine'), but also by its *place* (usually a geographical entity), and the *entities* (e.g, the Red Cross) involved in its reporting or resolution. Thus, identifying sentiment towards a frame involves consideration of all the above attributes.
- The sentiment expressed towards a given situation is not necessarily exclusive; both positive or negative sentiment can be expressed relating to different textual segments of the frame or even towards the same segment. Thus, the strict evaluation of the performance of the situation frame task involves identifying both the sentiment expressed as well as the associated segment of text expressing the sentiment.
- A situation frame is not anchored to a specific sentence or contiguous paragraph in the text; the same frame can be referred to in multiple parts of a document, and different frames can be referenced in the same sentence or paragraph. This requires a separate process, independent of the sentiment identification system,

to anchor the frame to a segment or segment(s) of text.

- The segments of text containing sentiment may or may not contain frame-relevant information. For example, the situation frame ‘infrastructure’ can be referred to in one sentence through the keywords ‘bridge’ or ‘demolished’, but the sentiment expressed towards the demolishment may not occur until several sentences later in the document.

The discussion in this section is only meant to be an introduction to the situation frame task along with possibilities for cross-lingual transfer, which we believe would be highly valuable for future applications that enable systems to quickly identify situations of urgent need in regions where a natural disaster or political incident occurs and facilitate an effective humanitarian response.

We first briefly describe and give examples of situation frames used in our data. In Section 6.3.2, we discuss frame anchoring. In Section 6.3.3, we describe an introductory baseline approach for identifying sentiment towards situation frames, with proposals for extending the approach using targeted sentiment models. We describe evaluation methodologies and present preliminary experiments and results for English and Spanish in Sections 6.3.4 and 6.3.5, and we conclude the chapter in Section 6.4.

6.3.1 Situation Frames

Situation frames are described by the following attributes:

- Place entity
- Need or Issue type
- Situation status (current or not current)
- Sentiment expressed towards the frame (positive and negative)
- Emotion expressed towards the frame (fear, joy, and anger)

If the situation is a need, the following attributes also characterize the frame:

- Resolution status of the need (resolved or not)
- Entity reporting the need
- Entity involved in the resolution of the need
- Urgency of the need

Thus, sentiment is one of several attributes that requires identification in order to garner a full picture of the situation manifested by the frame.

6.3.1.1 Data

Our situation frame data uses the text annotated by the Linguistic Data Consortium (LDC) as part of the DARPA Low Resource Languages for Emergent Incidents (LORELEI) low resource language program (Christianson et al., 2018). Tables 6.15 and 6.16 shows statistics of the data used in the 2019 Pilot Evaluation of Sentiment, Emotion, and Cognitive State (SEC) identification of the frames[9][10]. This data consists of newswire, discussion forums, and Twitter. Note that the number of sentiment annotations can exceed the number of frames, because sentiment is annotated at the segment level.

	# Documents	# Segments	# Frames	# Pos	#Neg
Train	76	968	324	57	269
Test	263	1253	611	26	327

Table 6.15: Statistics of English situation frame data with segment-level sentiment annotations (Pos: Positive Segments, Neg: Negative Segments).

Table 6.17 shows examples of the frame text and annotated segments expressing sentiment. The need and issue types are listed in Tables 6.18 and Tables 6.19. As can be seen from these examples, segments containing sentiment may or may not contain

[9] LDC2018E79_LORELEI_2019_SEC_Pilot_Training_Data_V1

[10] LDC2019E02_LORELEI_2019_SEC_Pilot_Evaluation_Data_V3.0

	# Documents	# Segments	# Frames	# Pos	#Neg
Train	48	392	146	13	141
Test	211	1123	240	19	352

Table 6.16: Statistics of Spanish situation frame data with segment-level annotations (Pos: Positive Segments, Neg: Negative Segments).

frame-relevant information (see Example 2: ‘Bad news.’) Thus, sentiment must be located in the surrounding context of the frame segments.

Example 1 (Frame: Shelter; Sentiment: Positive)
Still, it rattled nerves, causing people to vow to step up their emergency preparations. It’s been 21 years since a ‘quake approaching this size has hit the LA area. And it wasn’t that big a ‘quake at all. Always a good idea to prepare for a natural disaster. **In Colorado where I live it’s a good idea to have a suvival kit in your car in case a blizzard forces you to spend the night on the side of the road.**
Example 2 (Frames: Shelter, Infrastructure; Sentiment: Negative)
Tornado outbreak: 3-2-2012. Major damage and some deaths in Southern Indiana. Reports are one small town, Marysville, was leveled and the death count is rising. Storms are also in Kentucky and Tennessee and Ohio. Reports of at least 6 dead so far in Indiana. They are saying it was an F4 and half a mile wide with multiple vortices. **Bad news.** I have family in Nashville TN so i’m worried. Tornadoes have killed at least 8 people in Southern Indiana ...
Example 3 (Frames: Search_Rescue; Sentiment: Negative, Positive)
China stood still and sirens wailed Monday to mourn tens of thousands of earthquake victims in the country’s deadliest natural disaster in a generation. Construction workers, shopkeepers and bureaucrats across the bustling nation of 1.3 billion people paused for three minutes at 2:28 p.m. (0628 GMT) – exactly one week after the magnitude 7.9 quake hit central China. Air-raid sirens and the horns of cars and buses sounded in memory of the estimated 50,000 dead. **Rescuers also briefly halted work in the disaster zone, where the hunt for survivors turned glum despite remarkable survival tales among thousands buried.**

Table 6.17: Examples of annotated sentiment segments (bold) with document context and associated frame types.

Need Frame Types
Infrastructure Evacuation Shelter Food_Supply Medical_Assistance Water_Supply Search_Rescue Utilities_Energy_Sanitation

Table 6.18: Need frame types.

Issue Frame Types
Terrorism Regime_Change Crime_Violence

Table 6.19: Issue frame types.

6.3.2 Frame Anchoring

Because frames are not anchored to a specific part of the document, a separate process from sentiment identification - or potentially a joint frame and sentiment identification process - is required to identify the segments of text that correspond to the situation.

In this introductory work, we separate frame anchoring from sentiment identification, and consider two ways by which frames are anchored to the text: gold anchoring and keyword anchoring.

6.3.2.1 Gold Anchoring

While frames are not anchored to text, the LDC annotation process includes a 'description' field which allows annotators to describe their reasons for creating a frame annotation. This description field often includes text from the sentence that led the annotator to make the decision to annotate a frame. If the description field is non-empty, therefore, a simple search script allows us to locate this frame segment using the description text. We call this process 'gold anchoring', because the text segment is provided by the frame annotator.

In our Situation Frame data, all training data and English frames contain description text, while only half of the Spanish test frames contain a non-empty description. For the remaining Spanish frames with empty description fields, we use keyword anchoring.

6.3.2.2 Keyword Anchoring

Keyword anchoring uses the frame identification system of Littell et al. (2018); Muis et al. (2018); Chaudhary et al. (2019). This system relies on a combination of unsupervised keyword identification mechanisms (e.g word embeddings, clustering, and TF-IDF) and supervised frame identification mechanisms (e.g adversarial convolution network) trained on natural disaster corpora (ReliefWeb[11] and CrisisNet[12]) and annotated by Littell et al. (2018). We use this system to identify both *segments* corresponding to the unanchored Spanish situation frames along with their *frame types* (Tables 6.18 and 6.19.)

Once these key frame segments are located, we generate neighboring segments in a window w of segments preceding and following the anchored segment. All the generated candidate segments are then passed to another system for sentiment identification (Section 6.3.3), as shown in Figure 6.11.

In future, a joint process may be investigated that makes use of attention-based targeted sentiment mechanisms (Liu and Zhang, 2017) or methods for joint frame and sentiment identification.

6.3.3 Identifying Sentiment towards Frames

Our baseline method uses the deep learning model from Chapter 4, Section 4.1 to identify sentiment expressed in anchored frame segments and generated neighboring segments. Segments predicted to have neutral sentiment are discarded and frame segments expressing positive and negative sentiment are returned. The model is trained in its monolingual supervised form with either pre-trained or updatable word embeddings, depending on the best configuration for the language (English or Spanish).

As a future direction, one possible extension of this approach would be to apply a targeted sentiment system, such as that described in Section 6.2, to identify sentiment towards frame keywords. However, it is not clear how well this approach would work because the expression of sentiment *towards* target entities (e.g '**the Syrian**

[11] https://reliefweb.int/

[12] http://crisis.net/

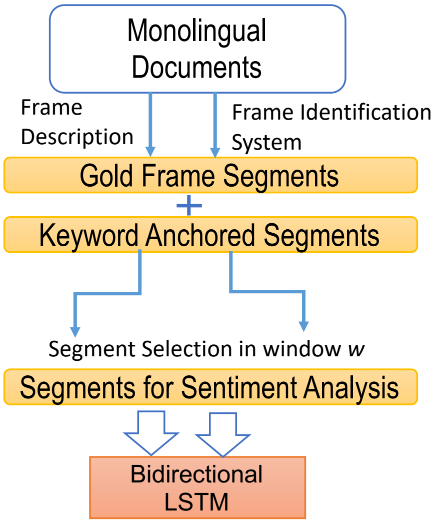

Monolingual Documents
Frame Description
Frame Identification System
Gold Frame Segments
+
Keyword Anchored Segments
Segment Selection in window *w*
Segments for Sentiment Analysis
Bidirectional LSTM

Figure 6.11: Anchoring and identifiying sentiment towards situation frames.

revolution continues') occurs differently than the expression of sentiment in relation to frame keywords (e.g 'Major **damage** and some deaths...multiple **vortices**. Bad news.') In the latter example, it is not clear that a model targeted to the 'infrastructure' keyword 'damage' or 'vortices' would capture the relationship between 'bad' and 'damage' or 'bad' and 'vortices' better than an untargeted model would. Example 3 from Table 6.17, however, is more straightforward in that it is easier to capture a relationship between the 'search_rescue' keyword 'hunt' and the sentiment word 'glum'.

6.3.4 Experiments

We present simple experiments with the goal of determining the baseline difficulty of the situation frame task by assessing the degree to which an untargeted sentiment model can capture sentiment expressed towards situations. The following factors are assessed:

- The baseline performance of the untargeted sentiment model at identifying targeted sentiment towards situations.

- The effect of the context window w when selecting candidate segments for sentiment analysis.
- The effect of gold anchoring vs. keyword anchoring for Spanish situation frames.

6.3.4.1 Data

The Situation Frame training datasets are quite small for the purposes of training a deep learning model, so we combined them with Twitter sentiment training data. To train the English untargeted sentiment model, we used a combination of the English Situation Frame training data and the SemEval 2015-2017 untargeted training data (Rosenthal et al., 2015b, 2017) described in Chapter 3. (The SemEval training dataset performed better than the European Twitter dataset for this task). Frame segments that have no sentiment annotated are labeled 'neutral' and combined with the Twitter data.

Similarly, to train the Spanish untargeted sentiment model, we used the European Twitter dataset (Chapter 3, Section 3.3) in combination with the Spanish Situation Frame training data). This training set is biased towards positive sentiment, in contrast to the test data, which is heavily biased towards negative sentiment, and so we downsampled the Spanish training data such that the sentiment distribution is 50% neutral, 25% positive and 25% negative.

6.3.4.2 Word Embeddings

We used pre-trained fixed GloVe 200-dimensional (Pennington et al., 2014) monolingual Twitter embeddings for the English model and randomly initialized updatable embeddings for the Spanish model, based on the configuration that worked best for the given language.

6.3.4.3 Metrics for Frame Sentiment Evaluation

We compute two metrics for the identification of sentiment towards frames and frame segments: a strict metric and a lenient metric that assigns partial credit if sentiment is correctly predicted towards the frame but the incorrect text segment is located.

A **strict tuple** is identified as a set: $\{document\_id, segment\_id, frame\_id, polarity\}$ where the first three are gold-assigned ids and $polarity \in \{positive, negative\}$.

A **partial tuple** is identified as a set: $\{document\_id, frame\_id, polarity\}$.

- **Strict Metric**: Computes the F-measure f_s of the number of matched strict tuples $matched_s$ between the gold frames and the model predictions.

$$r_s = \frac{matched_s}{gold_s} \tag{6.4}$$

$$p_s = \frac{matched_s}{predicted_s} \tag{6.5}$$

$$f_s = \frac{2r_s p_s}{r_s + p_s} \tag{6.6}$$

- **Partial Metric**: Computes the F-measure f_p of the number of matched partial tuples $matched_p$ between the gold frames and the model predictions.

$$r_p = \frac{matched_p}{gold_p} \tag{6.7}$$

$$p_p = \frac{matched_p}{predicted_p} \tag{6.8}$$

$$f_p = \frac{2r_p p_p}{r_p + p_p} \tag{6.9}$$

In addition, we compute the positive and negative class F-measures f_{pos_s}, f_{neg_s} and f_{pos_p}, f_{neg_p} for both strict and partial tuples.

6.3.5 Results

Table 6.20 shows the results. For English, where all frames are gold anchored, the untargeted model achieves a best strict score f_s of 0.40 without expanding the neighboring segment window and a partial score f_p of 0.53 when generating candidate segments in a window $w = 1$ around the gold anchored segment. Increasing the window size helps improve the partial score of identifying sentiment towards the frame, but not the strict score, which requires locating the exact text segment and therefore leads to a drop in precision when generating extra candidate segments.

Frame	Strict					Partial				
Anchoring	p_s	r_s	f_s	f_{pos_s}	f_{neg_s}	p_p	r_p	f_p	f_{pos_p}	f_{neg_p}
	English									
w=0	**0.40**	0.40	**0.40**	**0.098**	**0.42**	0.42	0.63	0.50	0.17	0.53
w=1	0.35	**0.43**	0.38	0.087	0.40	0.42	**0.74**	**0.53**	**0.20**	**0.56**
	Spanish									
w=0 gold	**0.11**	0.037	0.055	**0.050**	0.056	**0.46**	0.26	0.33	**0.39**	0.32
w=1 gold	0.10	0.067	0.080	0.030	0.088	0.44	0.42	**0.43**	0.33	0.45
w=0 kw	0.089	0.070	0.078	0.0	0.090	0.41	0.34	0.37	0.17	0.41
w=1 kw	0.069	0.12	0.088	0.015	0.12	0.25	0.42	0.32	0.26	0.45
w=0 gold+kw	0.094	0.092	0.093	0.025	0.100	0.42	0.40	0.41	0.29	0.43
w=1 gold+kw	0.075	**0.16**	**0.10**	0.025	**0.13**	0.26	**0.51**	0.35	0.27	**0.49**

Table 6.20: Baseline performance on detecting sentiment towards situation frames with strict($_s$) and partial($_p$) evaluation (r: recall, p: precision, f: F-measure, pos: positive class score, neg: negative class score, gold: gold frame anchoring, kw: keyword frame anchoring).

On the other hand, for Spanish, where only half the frames are gold anchored, increasing the window size helps to increase the strict score (from 0.055 to 0.08) by improving the recall of segments with identified sentiment (from 0.037 to 0.067). The partial score improves as well (from 0.33 to 0.43).

Similarly, using a keyword-based frame identification system helps to improve the performance of detecting sentiment towards Spanish situation frames (strict score increase of 0.055 to 0.078 for $w = 0$ and 0.080 to 0.088 for $w = 1$) by compensating for the missed frames that were not anchored in the text. However, this also leads to some drop in precision causing the partial score to decrease.

Combining both gold and keyword-based anchored frames with $w = 1$ leads to the best strict (0.1) and partial (0.49) scores for Spanish, resulting from improved recall of segments that particularly benefits the identification of negative sentiment (the dominant class), although positive class performance is reduced (0.025 strict score and 0.35 partial score). Therefore, if gold frames are not available or only partially available, using a keyword-based frame identification system along with window candidate segment generation is recommended.

Generally speaking, the performance of situation frame identification for Spanish is quite low compared to English, and this results from a combination of the bias of the Spanish evaluation dataset towards negative sentiment (352 negative sentiment annotations vs. only 13 positive sentiment annotations) as well as a lack of gold annotated frames which impacts frame segment recall in the strict score.

We were also able to use our bilingual English-Spanish embeddings (Chapter 4, Section 4.14) to improve the performance of frame sentiment detection for Spanish when using keyword anchoring, without any gold frame anchoring. We retained the same bilingual English-Spanish embeddings described in Chapter 4 and used them in the Spanish-trained model. The bilingual sentiment embeddings outpeformed the standard bilingual embeddings for $w = 1$ when using keyword anchoring, leading to a strict score of 0.1 using only keyword anchoring and a strict score of 0.12 using combined gold and keyword anchoring. However, the sentiment embeddings led to a drop in performance for other configurations ($w = 0$ and gold anchoring).

From this introductory exploration, there are several directions worth exploring in the future. One would be to train joint keyword-based frame and sentiment identification systems, in the spirit of the joint sentiment and target identification models of Yang and Cardie (2013), for example, that would help the keyword identification system better zone in on potential frame-relevant segments with sentiment content. Another would be to use a global targeted sentiment model, such as the CRF in Section 6.2, or a BiLSTM-CRF (Huang et al., 2015) to identify important frame keywords and the sentiment expressed towards them, where the sequence output would be aggregated by giving weights to different keywords and producing a single prediction for the frame. A third option would be to use an attention-based targeted model (such as one that will be referenced in Chapter 7) to identify sentiment towards specific frame keywords. Sentiment embeddings can be used in all these models and especially if the model were to be applied cross-lingually.

6.4 Conclusion

This chapter addressed the detection of sentiment towards targets in short documents, where the 'target' of sentiment is not straightforward to define or identify.

We first developed a two-stage method for defining and annotating targets of sentiment using the crowdsourcing platform Amazon Mechanical Turk, where targets are considered to be nominal entities. This method was applied to Arabic, yielding a new, publicly available resource of target-annotated comments to news articles for fine-grained entity sentiment analysis, the first resource of its kind for Arabic. We found high agreement on the task of identifying sentiment towards entities, leading to the conclusion that it is possible to carry out this task using crowdsourcing, especially when qualified workers are available.

Unlike some of the previous work, our focus was on annotating target entities rather the full target spans or only named entities; and we developed a unique approach for identifying these entities using Amazon Mechanial Turk. The first task involves marking important entities, while the second task involves finding targets by assessing the sentiment towards each entity in isolation. We found that although the agreement was generally high for both tasks, it was not as high for the entity identification task as it was for the second and easier task of finding sentiment towards entities. We also found that the morphological complexity of Arabic, as well as the variation in acceptable syntax for noun phrases, creates additional annotation challenges for deciphering the boundaries of entities. The long, complex, and often informal structure of the comments creates interesting challenges for modeling tasks.

This dataset was then used to develop linguistically inspired sequence labeling models that identify important entity targets along with sentiment in the news comments. The sequence labeling models are run in pipeline fashion and can operate under high-resource scenarios (with rich linguistic features, such as syntax) or more low-resource scenarios (using only part-of-speech tags and word embedding clusters). Both target and sentiment results significantly improve multiple lexical baselines and are comparable to previously published results in similar tasks for English, a similarly hard task. We showed that the choice of morphological representation significantly affects the performance of the target and sentiment models, as it does with transferring sentiment from Arabic. This could shed light on further research in target-specific sentiment analysis for morphologically complex languages, an area little investigated previously.

Finally, we introduced a new targeted sentiment analysis task: the identification of sentiment towards situations, which would be valuable for providing humanitarian assistance to locals in need during natural disasters or indicents of violence. Our baseline system performed reasonably well when using gold and keyword anchoring of situation frame text, noting differences in performance between English and Spanish, where gold frame anchoring was not fully available. By combining gold anchored frames with keyword anchoring and a larger window for segment generation, improvements to the Spanish system were made possible. Our preliminary results show the promise of this new application with future monolingual and cross-lingual models.

Chapter 7

Transferring Targeted Sentiment Cross-lingually

The end goal of a sentiment application that can respond to the sentiments and needs of locals speaking a low-resource language is to identify sentiment towards targets cross-lingually. Transferring targeted sentiment cross-lingually requires training a model that can identify sentiment expressed towards targets in a high-resource language, and then applying it to the low-resource language. This is a problem that has been investigated very little in natural language processing, but that would be highly valuable not only for targeted sentiment analysis in low resource languages, but also in many high-resource languages, where datasets annotated for targeted sentiment are much more scarce compared to untargeted sentiment.

This final chapter builds on all the methods and resources presented in the dissertation, drawing on bilingual features, resources, language, and target, in order to predict sentiment towards targets cross-lingually. Because the problem is a relatively new one and its difficulty has yet to be assessed, the targets considered in this chapter are usually entities that occur in shorter text than that considered in the previous chapter, and the genre of our evaluation text is instead more similar to that described in Chapters 4 and 5. The methods presented in this chapter can be considered as a foundation that can be used to develop targeted cross-lingual models in text where targets are less explicitly defined, such as situation frames.

We adapt our cross-lingual model, presented in Chapter 4, to the prediction of sentiment towards candidate targets, which are provided as input to the model. This targeted cross-lingual model makes use of the bilingual feature representations and resources that we have developed and presented through this work: specifically, bilingual embeddings trained on translation corpora, target language lexicalization and updatable bilingual sentiment embeddings, morphological representations, and both in-domain and out-of-genre translation corpora. Throughout, we investigate whether the techniques that worked well for untargeted transfer hold for the transfer of targeted sentiment. We find that our conclusions for untargeted sentiment as to the effect of in-domain corpora, bilingual feature representations and morphological representations, in the majority of cases hold true for targeted sentiment, in experiments that we present for English to Arabic and Arabic to English transfer. In particular, our bilingual feature representations created by incorporating sentiment embeddings and updatable weights with lexicalization, result in notable improvements to identifying sentiment towards targets cross-lingually.

Section 7.1 describes our cross-lingual sentiment model, adapted for identifying sentiment towards targets. In Section 7.2, we describe the bilingual feature representations and resources used for transfer of targeted sentiment. Section 7.2.4 discusses the morphological representations used for transferring sentiment towards targets with Arabic as a source language. Section 7.3 describes experiments, where we show the performance of the targeted cross-lingual sentiment model on two-class (positive and negative) and three-class (positive, negative, and neutral) sentiment identification. Section 7.4 presents our results, Section 7.5 presents an analysis of errors in the transfer of targeted sentiment, and we conclude in Section 7.6.

7.1 Model Architecture for Cross-lingual Transfer of Targeted Sentiment

To identify sentiment towards targets, we incorporate an *attention* mechanism to the bidirectional long short-term memory (biLSTM) model. Applying attention to biL-

STM was first introduced by Bahdanau et al. (2014) for neural machine translation, and has since been used by many natural language processing tasks, such as parsing, document classification, and sentiment analysis - including Liu and Zhang (2017) who used attention to model targeted sentiment using an LSTM, and Wang et al. (2016), who used attention modeling in a similar manner, except that it was applied towards aspects rather than targets. In this work, we apply attention modeling to targets in a cross-lingual sentiment model.

The input to our cross-lingual targeted model is a sequence of words, $x = \{x_1, x_2, \cdots, x_n\}$, and a target $t = \{t_1, t_2, \cdots, t_m\}$, $m<n$, where n is the length of the sequence and m is the length of the target entity. This input is fed into the bidirectional layer, as with our untargeted cross-lingual model. In a standard LSTM or biLSTM, the hidden layer output h_n - or forward and backward outputs ${h_n}^{forward}$ and ${h_n}^{backward}$ if the model is bidirectional - at the last timestep n is passed to the next layer or output layer for computing probabilities for each class label (as shown in Figure 4.1). This last hidden layer output h_n therefore represents the meaning or context of the sentence. With attention modeling, all the hidden states $h_1, h_2, \cdots h_n$ contribute to the representation of the sentence by assigning an attention weight a_i to the hidden state at each timestep. This allows the model to 'pay attention' to specific words in the sentence, such as sentiment-bearing words, that may contribute more heavily to the prediction of the sentiment label. When modeling attention towards the *target*, the model is allowed to pay attention to words that contribute to the sentiment expression *as well as* the target.

With attention modeling, the representation of the sentence s becomes:

$$s = \sum_{i=1}^{n} \alpha_i h_i, \tag{7.1}$$

The weights a_i for each hidden state h_i are computed by a 'softmax' probability distribution as follows:

$$\alpha_i = \frac{exp(\beta_i)}{\sum_{j=1}^{n} exp(\beta_j)}. \tag{7.2}$$

To compute β for targeted attention, we use the contextualized attention model of Liu and Zhang (2017):

$$\beta_i = f(h_i; h_t), \tag{7.3}$$

where f is a *tanh* feedforward network that is applied to the concatenation of the hidden state representation h_i and the target hidden state representation h_t, and in our case h_t is computed by summing the hidden state vectors for all target words.

Moreover, in the contextualized attention representation of Liu and Zhang (2017), the input sequence is divided into a left sequence $h_1, h_2 \cdots h_{t_1-1}$ (words preceding the target) and 'right sequence' $h_{t_m+1} \cdots h_n$ (words following the target), and attention is computed separately for the left and right contexts. Computing sentiment in left and right contexts has proved useful for targeted models in previous studies (e.g, that of Tang et al. (2015)). Note that attention is not applied to the target words themselves.

$$s_l = attention([h_1 \cdots, h_{t_1-1}]; h_t), \tag{7.4}$$

$$s_r = attention([h_{t_m+1} \cdots, h_n]; h_t), \tag{7.5}$$

As our cross-lingual model is bidirectional, a hidden state h_j is computed by concatenating the forward hidden state ${h_j}^{forward}$ and ${h_j}^{backward}$, as in the original work that proposed attention modeling to create soft word alignments for machine translation (Bahdanau et al., 2014).

Our model concatenates the full sentence representation s, left sequence representation s_l, and right sequence representation s_r along with our average pooling layer

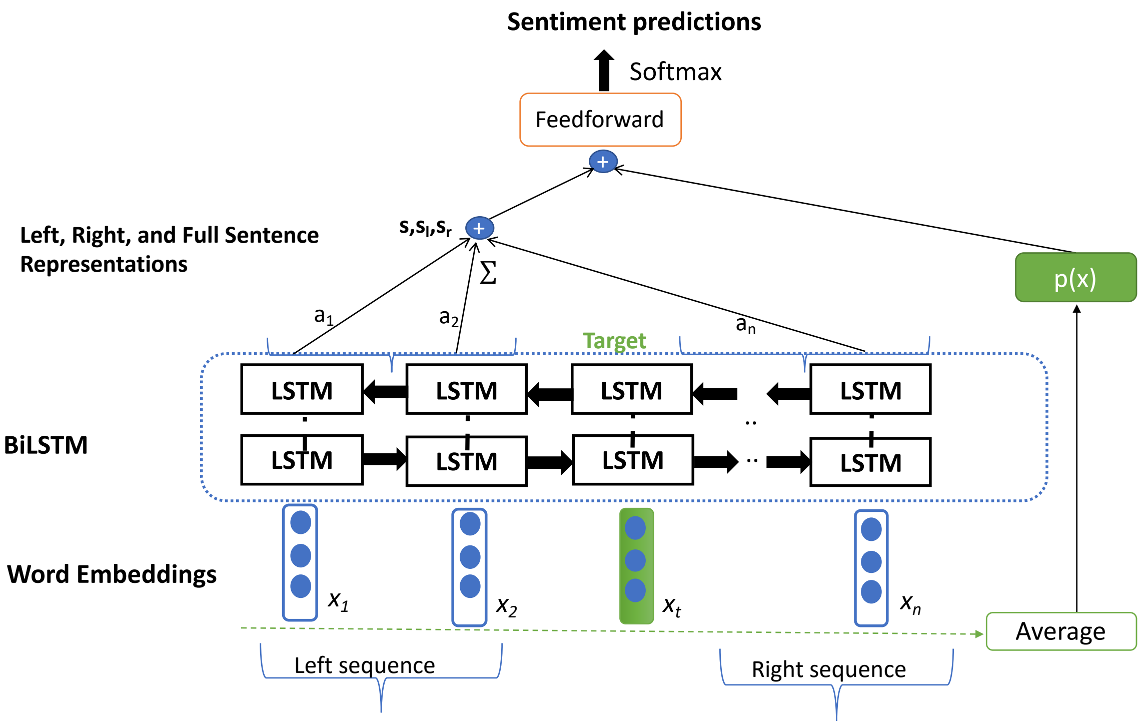


Figure 7.1: Model architecture for cross-lingual transfer of targeted sentiment. Attention weights $a_1, a_2 \cdots a_n$ are computed for each biLSTM hidden state and are used to compute left, right, and full context sentence representations s_l, s_r, and s.

p, before passing the result through a feedforward layer with 'softmax' activation that computes conditional probabilities for each sentiment label. A diagram of the architecture is shown in Figure 7.1.

In the context of the cross-lingual model, the attention model not only attends to the target, but also helps us attend to source language words, or target language words if lexicalization is applied, that contribute to the identification of sentiment. Therefore, when bilingual sentiment embeddings and target language lexicalization are applied, the attention mechanism can attend to the bilingual word representations that most contribute to the identification of sentiment towards the target.

7.2 Bilingual Feature Representations and Resources

We describe the bilingual features and representations that led to successful cross-lingual transfer of untargeted sentiment, as shown in Chapter 4, as well as the morphological representations that led to successful cross-lingual transfer of untargeted sentiment and identification of targeted sentiment, as shown in Chapters 5 and 6, and that we now integrate into the targeted cross-lingual sentiment model.

7.2.1 Bilingual-based Embeddings

The targeted cross-lingual model uses pre-trained cross-lingual word embeddings (BL) trained on parallel translation corpora (described in Chapter 4, Section 4.2.1). Embeddings trained on translation corpora, even if relatively small, were determined to result in the best cross-lingual sentiment performance in the majority of languages when identifying untargeted sentiment, and we therefore use these pre-trained embeddings in our targeted cross-lingual model.

7.2.2 Bilingual Sentiment Embeddings and Lexicalization

In addition to pre-trained bilingual-based embeddings, we create additional bilingual feature representations for the targeted cross-lingual transfer model (described in Chapter 4, Section 4.2.2). We use the feature configuration that worked best overall for most languages, which corresponds to the 'BSW+Lex' representation features. This configuration uses target language lexicalization - partial translation of the training data into the target language using a bilingual dictionary created from the parallel corpus, along with pre-trained bilingual sentiment embeddings with weight update, and bilingual sentiment scores $v_{sentiment}$ with sentiment weight update. Target lan-

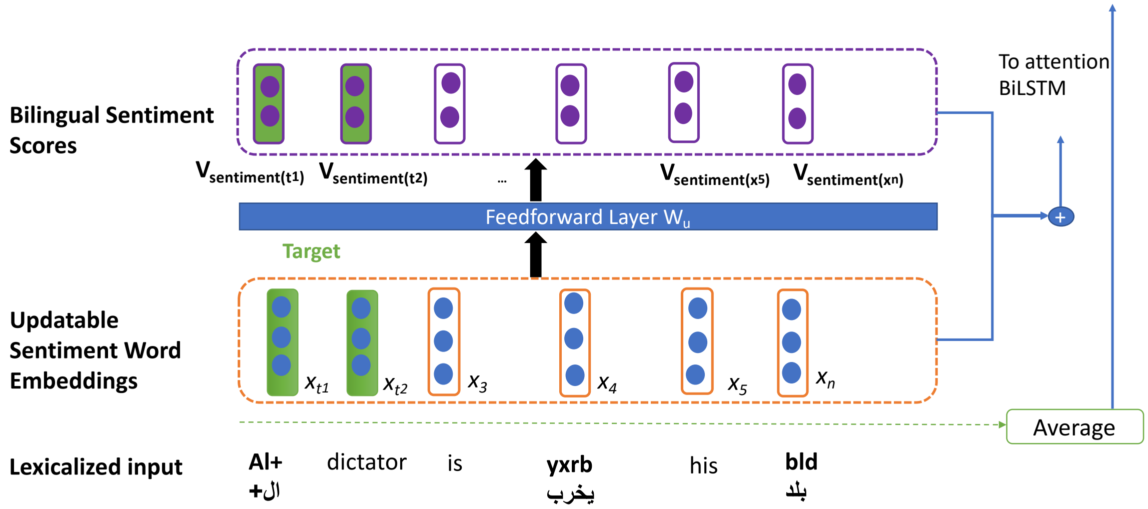


Figure 7.2: Model architecture for cross-lingual transfer of targeted sentiment, with updatable bilingual sentiment embeddings and weights, and lexicalized input. Sentiment embeddings and scores are passed to the BiLSTM attention model. The English sentence 'the dictator is destroying his country' is partially lexicalized to Arabic.

guage lexicalization allows both sentiment embeddings and output sentiment weights to be updated during training.

With the targeted cross-lingual model, the integration of the attention mechanism means that the model has the opportunity to give weight to the representations of different words in the input, and in the case of lexicalization during training, this would include both source-language and target-language words, when they are found in the bilingual dictionary.

Figure 7.2 shows how bilingual sentiment weight and target language lexicalization representation features are created before being passed on to the attention biLSTM.

7.2.3 In-domain and Out-of-domain Corpora

To pre-train bilingual-based word embeddings and sentiment embeddings, two translation resources are used: in-genre and in-domain parallel corpora from the Linguistic Data Consortium (LDC) and out-of-domain parallel corpora from the Bible and Quran (QB).

Previously, we showed that in-domain parallel corpora generally result in the best configuration of cross-lingual transfer of untargeted sentiment, even when the in-domain corpus is of relatively smaller size. However, out-of-domain corpora were shown to be a viable alternative when in-domain corpora are not available. For our targeted cross-lingual sentiment model, we include feature representations created from both of these resources.

7.2.4 Morphological Representations

In this part of the work, Arabic is used as both a target language and a source language. Throughout the thesis, it has been shown that the choice of morphological representations impacts the transfer of sentiment from Arabic as well as the identification of sentiment towards targets in Arabic. In this part, we consider the two morphological segmentation schemes, ATB and D3, which were shown to reduce the sparsity of the vocabulary by splitting off clitics: all clitics in the case of D3, and all clitics except the definite article in the case of ATB. These schemes are described in Chapter 6 and are used for transferring targeted sentiment from Arabic.

When transferring targeted sentiment to Arabic from English, we use the same heuristic form-based ATB scheme used in Chapter 4 in untargeted cross-lingual experiments with Arabic as a target language. All parallel corpora (LDC, QB), training datasets, and evaluation datasets for Arabic are preprocessed accordingly before training bilingual word embeddings in English-to-Arabic and Arabic-to-English transfer configurations.

While *lemma-based* representations were found helpful for identifying sentiment towards targets in Chapter 6, we keep the surface word representations in this set of experiments, for consistency with the untargeted cross-lingual experiments in earlier chapters. However, we expect that using lemma-based bilingual word embeddings and datasets would be helpful for cross-lingual and targeted cross-lingual sentiment analysis in Arabic.

7.3 Experiments

The goal of the experiments in this chapter is to assess the performance of the targeted cross-lingual sentiment model and to determine whether the settings that worked best for transferring untargeted sentiment hold for the the transfer of sentiment towards targets. Therefore, the attention-based targeted model with our best feature configuration will be compared with the attention-based targeted model inspired from Liu and Zhang (2017) that relies only on attention modeling, and in-domain corpora will be compared with out-of-genre corpora.

Our experiments are run on two languages: English and Arabic, and we evaluate the performance of the cross-lingual model when it is trained on English and tested on Arabic, as well as when it is trained on Arabic and tested on English.

Moreover, because of the greater difficulty assumed to be involved in the task of transferring the sentiment towards targets - syntactic relationships may not hold across languages, for example - we evaluate the model in two settings: the easier *two-class* prediction, where targets are known, but it is required to identify whether sentiment is positive or negative, and *three-class* prediction, where it is required to identify both whether sentiment is expressed towards the candidate target as well as the polarity (positive, negative, and neutral). The two-class prediction also allows us to compare the performance of the cross-lingual model with the supervised Arabic results obtained using our open-domain dataset.

The targeted cross-lingual experiments, therefore, assess the following factors:

1. The performance of the attention-based targeted cross-lingual model, compared to the untargeted cross-lingual model, when identifying sentiment towards targets cross-lingually.

2. The effect of target language lexicalization and bilingual sentiment embeddings with weight update, on the identification of sentiment towards targets using the attention-based targeted model.
3. The effect of bilingual resources, in-domain and out-of-domain parallel corpora, on the identification of sentiment towards targets.
4. The performance of different morphological segmentation schemes, when identifying sentiment towards targets cross-lingually and using Arabic as a source language.

In what follows, we describe experimental setup and configurations, including the data used for training and evaluating the targeted cross-lingual sentiment model.

7.3.1 Setup and Configurations

We adapted our cross-lingual sentiment model by implementing an attention mechanism towards input targets as described in Section 7.1. The code for the targeted model is publicly available[1].

To train bilingual word and sentiment embeddings, we used the same configurations, with our update of the Multivec (Bérard et al., 2016) toolkit, as described in Chapters 4 and 5. New English-Arabic bilingual embeddings were trained after processing the LDC and QB corpora using the high-resource ATB tokenization scheme. For lexicalization, we used the bilingual dictionaries that were created accordingly using the specified parallel corpora.

The targeted cross-lingual model was trained using the same parameters described for the bilingual-based model in Chapter 4: 5 training epochs, batch size of 32, and categorical cross entropy training with optimization using the Adam (Kingma and Ba, 2014) optimizer.

[1] https://github.com/narnoura/cross-lingual-sentiment

7.3.2 Data

To train and test the targeted cross-lingual models, we used the targeted SemEval-2017 (Rosenthal et al., 2017) datasets for SemEval Subtasks B,D and C,E[2], and the Dong Twitter (Dong et al., 2014) datasets, both described in Chapter 3, Section 3.3.2. Both datasets consist of Twitter data, and are thus of the same genre as the untargeted datasets used for training and evaluating the untargeted English and Arabic cross-lingual models in earlier chapters. Each data sample is annotated for sentiment towards a topic (the target), where for the SemEval data, the topic is the keyword that was used to collect the tweet. Examples of these tweets are provided in Tables 3.8 and 3.9.

7.3.2.1 Two-Class Prediction

Our two-class prediction models use the SemEval Sentiment in Twitter Subtask B,D data, which involves identifying positive or negative sentiment towards topics. For Arabic, we used the full SemEval 2017 Task B,D training and test sets. For English, which contains no new training data from 2017, we used the 2016 train and test data for training, and the 2017 Task B,D test data for evaluation. We set aside the 2016 dev and devtest datasets for development, and we did not use the 2015 B,D training data, because it contains neutral labels.

7.3.2.2 Three-Class Prediction

For Arabic three-class prediction, we used the new Arabic data we collected for the SemEval Sentiment in Twitter Subtasks C,E, which involve identifying sentiment on a five-point scale towards targets, and which we consolidated to a three-point scale as described in Chapter 3.

[2] Tasks B and D involve identification of targeted sentiment on a two-point scale in Twitter, while Tasks C and E involve identification of targeted sentiment on a five-point scale.

For English three-class prediction, we used the Dong Twitter benchmark dataset (Dong et al., 2014) for training and evaluation[3].

Table 7.1 shows the size and distribution of the SemEval datasets for the data used in the experiments. Neutral (O) labels are excluded for 2-class experiments. For English, we use the Dong dataset for 3-class experiments, which is distributed evenly among neutral (50%), positive (25%), and negative (25%) labels for train and test.

SemEval Datasets	English		Arabic	
Train	14,897		3355	
% P	11803	(79.2%)	885	(26.4%)
% N	3095	(20.8%)	771	(23.0%)
% O	–	–	1699	(50.6%)
Test	6,185		6100	
% P	2463	(39.8%)	1561	(25.6%)
% N	3722	(60.2%)	1196	(19.6%)
% O	–	–	3343	(54.8%)

Table 7.1: SemEval 2017 English and Arabic targeted sentiment datasets with train and test sizes, number of samples, and distribution in percentage % amongst sentiment labels (P:positive, N:negative, O:neutral).

7.3.3 Metric

As with untargeted cross-lingual sentiment evaluation, we use macro-averaged F-measure (F-Macro), averaged over the two or three sentiment classes, as the main metric for evaluating the performance of the sentiment predicted towards targets.

We additionally computed accuracy, to provide a fuller picture of the performance of the cross-lingual model - for example, if Accuracy is high but F-Macro is low, this is a sign that the model often predicts the majority class.

Because of the small size of some of the training datasets and the resulting fluctuation in results, we computed the mean of metrics averaged over 10 runs for each

[3] We noticed that the README file for the SemEval-2017 task C,E training set, which was collected previously to 2017, indicated that the '0' label included 'negative or neutral' labels, while our task requires that it be mapped only to neutral labels, so we used the Dong dataset instead.

experiment. The confidence intervals we computed across the different results fall in the range of 0.3 point intervals to 2.4 point intervals (for Accuracy) and 0.6 point intervals to 2.9 point intervals (for F-Macro).

7.4 Results

Table 7.2 shows the results for identifying sentiment towards targets in Arabic, and 7.3 shows the results for identifying sentiment towards targets in English.

Supervised Systems. The first row shows the supervised in-language baseline, which was run using randomly initialized and updatable embeddings from the training data. For English, the best supervised model among targeted and untargeted models was selected. For Arabic, the best supervised model among targeted and untargeted models, and D3 and ATB tokenization schemes was selected. We note that because the training datasets are relatively small, using pre-trained monolingual embeddings from a larger corpus would increase these numbers; for example, using GloVe (Pennington et al., 2014) English Twitter embeddings with the targeted English model resulted in an accuracy of 68.2 and macro-averaged F-measure of 66.6 on the three-class Dong test set.

Majority Baselines. The next row shows the performance of a model that always predicts the MAJORITY baseline of the source language (positive for 2-class and neutral for 3-class, for both English and Arabic training sets) on the target language test data.

Cross-lingual Models. The next three rows show the cross-lingual models used for identifying sentiment towards targets. The *Untargeted* model refers to our untargeted cross-lingual model from Chapter 4, which is used as is for predicting sentiment towards targets without considering information related to the target. The *Targeted* model refers to the attention-based targeted cross-lingual model described in Section

7.1. Finally, *+BSW+Lex* refers to the attention-based targeted cross-lingual model that uses our feature configuration of target language lexicalization and bilingual sentiment embeddings and weights.

The bilingual features are created using the LDC (Linguistic Data Consortium) or Quran-Bible (QB) parallel corpora respectively. For transferring from Arabic to English, we show results with the two preprocessing schemes: ATB (full tokenization except for the definite article) and D3 (full tokenization).

Method	Accuracy	F-Macro
2-class (en-ar)		
Supervised(ar)	68.4	67.2
Majority	56.7	36.2
Untargeted *(LDC)*	**73.6**	**72.5**
Targeted *(LDC)*	73.4	72.4
+BSW+Lex *(LDC)*	72.6	72.2
Untargeted *(QB)*	**63.4**	**60.4**
Targeted *(QB)*	62.8	59.5
+BSW+Lex *(QB)*	60.9	57.7
3-class (en-ar)		
Supervised(ar)	49.7	42.7
Majority	54.8	23.6
Untargeted *(LDC)*	52.8	38.3
Targeted *(LDC)*	**53.1**	39.6
+BSW+Lex *(LDC)*	51.3	**43.3**
Untargeted *(QB)*	44.3	34.2
Targeted *(QB)*	**45.7**	34.6
+BSW+Lex *(QB)*	42.3	**35.0**

Table 7.2: Accuracy and Macro-averaged F-measure (F-Macro) for predicting 2-class and 3-class targeted sentiment in Arabic using Supervised model, Majority baseline of source language, and Cross-lingual targeted models (Untargeted: no attention to target, Targeted: attention-based model, +BSW+Lex: attention-based model with bilingual sentiment embeddings and weights).

Method	Accuracy		F-Macro	
2-class (ar-en)				
Supervised(en)	68.7		68.7	
Majority	39.8		28.5	
	D3	ATB	D3	ATB
Untargeted *(LDC)*	55.3	56.3	53.9	55.1
Targeted *(LDC)*	58.1	57.4	57.3	56.4
+BSW+Lex *(LDC)*	62.5	**65.0**	62.3	**64.8**
Untargeted *(QB)*	58.1	60.2	57.8	**60.0**
Targeted *(QB)*	57.3	**60.5**	56.8	**60.0**
+BSW+Lex *(QB)*	55.0	58.3	54.5	58.2
3-class (ar-en)				
Supervised(en)	63.1		61.3	
Majority	50.0		22.2	
	D3	ATB	D3	ATB
Untargeted *(LDC)*	43.7	42.5	36.0	36.4
Targeted *(LDC)*	42.0	42.1	36.6	36.3
+BSW+Lex *(LDC)*	**44.7**	43.6	38.9	**40.2**
Untargeted *(QB)*	43.7	45.8	34.7	34.6
Targeted *(QB)*	45.8	45.6	34.3	35.2
+BSW+Lex *(QB)*	45.1	**47.2**	36.3	**37.7**

Table 7.3: Accuracy and Macro-averaged F-measure (F-Macro) for predicting 2-class and 3-class targeted sentiment in English using Supervised model, Majority baseline of source language, and Cross-lingual targeted models (Untargeted: no attention, Targeted: attention mechanism, +BSW+Lex: attention mechanism with bilingual sentiment embeddings and weights.)

7.4.1 Evaluation of Targeted Transfer Model

First, we can see that for two-class and three-class sentiment prediction, all cross-lingual models outperform the majority baseline in terms of the main evaluation metric, F-Macro. Moreover, when transferring from English to Arabic, the best cross-lingual model (untargeted with LDC for 2-class and BSW+Lex with LDC for 3-class) surpasses the supervised Arabic model. This is because the supervised model uses only the small Arabic training data, while the cross-lingual model uses pre-trained embeddings from bilingual corpora in addition to the English training data. This result is promising, as it means that cross-lingual targeted transfer models can be used as a means of providing or increasing training data for languages which do not have very large targeted sentiment annotation datasets.

When considering the impact of the targeted attention mechanism on the performance of the cross-lingual model, the results are somewhat mixed. The target attention mechanism helps most when transferring from Arabic to English, and in 3-class prediction when transferring from English to Arabic. In some cases, there is no significant difference in results between the attention-based targeted model and the untargeted model. For example, in 2-class sentiment transfer from English to Arabic (Table 7.2), all cross-lingual models perform relatively similarly. The confidence intervals for F-Macro with LDC are 72.5 ± 0.51 for the untargeted model and 72.4 ± 0.64 for the attention-based model, and the differences are not significant. Similarly for QB, the confidence intervals are 60.4 ± 1.4 for the untargeted model and 59.5 ± 1.9 for the attention-based model, indicating a large overlap region and no significant difference. The differences are more apparent for the harder 3-class prediction of sentiment towards targets (38.3 ± 1.5 for untargeted and 39.6 ± 0.82 for targeted attention), as well as English to Arabic transfer (F-Macro of 53.9 vs. 57.3 with D3 and 55.1 vs. 56.4 with ATB for 2-class prediction in Table 7.3). The attention mechanism also helps the performance of positive and negative class prediction, as our error analysis shows. However, the degree of improvements gained by the contextualized attention model are not far in line from the supervised English results reported by (Liu and Zhang, 2017), which was about an increase in 1 point in F-Macro and Accuracy on average. This was also consistent with our results when we evaluated our SUPERVISED model on English: 61.3 Accuracy and 59.3 F-Macro on three class prediction without attention, and 63.1 Accuracy and 61.3 F-Macro with targeted attention.

In most cases, the addition of target lexicalization and bilingual sentiment embedding weight features to the cross-lingual targeted attention model gives it a boost, as we detail in the following section.

7.4.2 Evaluation of Bilingual Feature Representations

From Tables 7.2 and 7.3, we can see that using target language lexicalization with updatable bilingual sentiment embeddings and weights results in the best targeted model configuration in all cases except 2-class sentiment prediction when transferring from English to Arabic, where it performs indistinguishably from the untargeted model (72.2 vs 72.5 F-Macro using LDC).

On the other hand, with the more difficult task of 3-class sentiment prediction towards targets, we observe that using BSW+Lex results in the best overall English-Arabic (Table 7.3) targeted transfer result (43.3±0.72 F-Macro with LDC), exceeding the supervised model result (42.7) and also leads to an improvement with QB (35.0). Similarly, with Arabic-English targeted transfer (Table 7.2), BSW+Lex results in the best overall result of 64.8±1.8 with ATB and 62.3±1.9 with D3 for 2-class prediction, coming quite close to the English-trained supervised model, and the best overall result of 40.2±1.2 with 3-class prediction using LDC. These results are consistent with what we observed for untargeted cross-lingual sentiment transfer in Chapter 6.

Moreover, BSW+Lex gives a boost to the performance of the attention mechanism, often outperforming the vanilla version of the targeted sentiment model. The improvement gained from using BSW+Lex is not as consistent with QB embeddings as it is with LDC embeddings, which is again in line with untargeted transfer results in Chapter 6.

7.4.3 Evaluation of Bilingual Feature Resources

As was expected, using bilingual features trained on the in-domain and in-genre LDC corpus results in the best cross-lingual performance in the majority of scenarios when transferring targeted sentiment from English to Arabic and Arabic to English. The only scenario where this is not the case is the performance of the Arabic-English 2-class vanilla prediction model (Table 7.3). In this case, QB embeddings lead to better performance for the untargeted model (57.8 vs. 53.9 with D3 and 60.0 vs. 55.1 with ATB) and the targeted model for ATB (60.0 vs. 56.4). It is unclear why the out-of-domain corpus fares better here - perhaps, since the QB corpus is larger than the LDC corpus, it makes up for the smaller Arabic training data; however, upon adding BSW+Lex with the targeted attention mechanism, LDC embeddings result in the best performance for both ATB and D3.

7.4.4 Evaluation of Morphological Schemes

Here we compare the performance of the D3 and ATB tokenization schemes when transferring targeted sentiment from Arabic to English. We observe some differences in sentiment prediction performance, in keeping with the results observed in Chapter 6, where it was found that training with ATB is better overall for identifying sentiment towards targets. The situation is a little bit different here, in that the English evaluation data does not contain split tokens. However, it does indicate that in most cases, splitting tokens at ATB level is compatible with English evaluation data, and splitting off the definite article when transferring from Arabic is not necessary, while keeping it likely introduces some more noise in the attention model that is less relevant to the prediction of sentiment.

Unlike what was observed in Chapter 6, there seems to be no advantage to using D3 for identifying targets with 3-class prediction in English evaluation data (or the advantage is over-ridden by the better performance of ATB on positive vs. negative sentiment prediction), even though the definite article is always separated in English. This might be because of differences in the way language is used - in Arabic, the definite article often indicates emphasis to important entities that is not necessarily resembled in English.

The results comparing the two schemes are also in line with those in Chapter 5 when transferring untargeted sentiment from Arabic to English.

7.4.4.1 Benchmark Comparison

On a final note, the results for identifying 2-class sentiment towards targets cross-lingually in Arabic are in line with our supervised Arabic targeted results in Chapter 6: 73.6 best model accuracy and 72.5 F-Macro using the cross-lingual model, compared to 76 best model Accuracy and 75 F-Macro using the supervised CRF model in Chapter 6 - even though the data is quite different. The second best-performing supervised Arabic targeted system in the SemEval 2017 task B competition (Rosenthal et al., 2017) achieved an accuracy of 73.4 and F-Macro of 72.1 on this task. The best system, which used a large amount of external Arabic training data, achieved an accuracy of 77 and F-Macro of 76.7. (Our supervised Arabic model does not perform as well, because we have used only the SemEval training data without any external embeddings.) This helps establish an expectation for the performance of targeted and cross-lingual targeted sentiment models in Arabic, for which future studies can compare to.

7.5 Error Analysis

We examined the F-measure breakdown of the untargeted, attention-based targeted, and attention-based targeted lexicalized bilingual sentiment models, as well as output samples where their predictions were different, to see where and whether they differed in identifying sentiment towards targets.

Upon examination of the positive and negative class F-measures of the English-to-Arabic two-class transfer model, where the attention-based targeted models did not outperform the untargeted model, we found that the three models still differed in behavior. In particular, whereas the overall accuracy and F-measure was similar amongst the three models, the lexicalized sentiment embedding model attempted to balance sentiment prediction among the two classes by predicting more negative sentiment when the positive class was the majority. Table 7.4 shows the breakdown of the 2-class and 3-class prediction F-measures, when using the LDC corpus. While the 2-class English SemEval training set is biased towards positive sentiment, the 3-class English Dong training set is balanced equally between positive and negative classes, with 50% neutral sentiment.

	Acc	F-Macro	F-Pos	F-Neg	F-Neut
			2-class		
Untargeted	**73.6**	**72.5**	**78.0**	67.0	–
Targeted	73.4	72.4	77.6	67.2	–
+BSW+Lex	72.6	72.2	75.4	**69.1**	–
			3-class		
Untargeted	52.8	38.3	25.5	22.4	67.0
Targeted	**53.1**	39.6	23.0	28.4	**67.3**
+BSW+Lex	51.3	**43.3**	**34.0**	**32.5**	63.4

Table 7.4: Accuracy, F-measure, and breakdown of F-measure for positive (F-Pos), negative (F-Neg), and neutral (F-Neut) classes for untargeted, targeted, and targeted cross-lingual models with BSW+Lex for English to Arabic cross-lingual prediction of sentiment towards targets. Results are averaged over 10 runs.

We can also see that while the targeted model with attention to targets improves the overall F-measure and negative F-measure of 3-class targeted sentiment prediction, incorporating the lexicalized and bilingual sentiment features along with attention to targets leads to a substantial increase in both positive and negative as well as overall prediction scores, without substantially decreasing the performance of the neutral class.

Tables 7.5 and 7.6 show examples of the output of the three cross-lingual models on English-to-Arabic and Arabic-to-English transfer of targeted sentiment. LDC feature representations are used for both models and the ATB scheme is used to preprocess the Arabic-trained model.

Input **(en-ar)**	Untargeted	Targeted	BSW+Lex	Gold
@user I want **Ramy Ayach**	neutral	**positive**	**positive**	positive
United Nations: Not enough capacity to treat the injured in Mosul — Iraq	negative	negative	**neutral**	neutral
And with him Lebanese journalism was assassinated... **Gebran Tueni**	neutral	**positive**	negative	positive
350 Palestinian children are sitting in prisons ..**the occupation**	neutral	neutral	**negative**	negative
Jaafari launches a heated attack against the United States embassador to **the United Nations**	**neutral**	negative	negative	neutral

Table 7.5: Example of output predictions with translated input for targeted cross-lingual sentiment model trained on English and evaluated on identifying sentiment towards targets in Arabic. The target is indicated in bold.

Input **(ar-en)**	Untargeted	Targeted	BSW+Lex	Gold
it's official: **george bush** was such a bad president that you can win the nobel peace prize just by not being him . fb	positive	neutral	**negative**	negative
i love you **britney spears** but i do not like your new song : / * changes channel	neutral	**positive**	**positive**	positive
I Cant Wait for **harry potter** and the half blood prince to come out on dvd december 7th !!!	neutral	positive	**neutral**	neutral
i like winamp, but since getting my **ipod** touch i use itunes, and it's growing on me.	**neutral**	positive	**neutral**	neutral
i heard ShannonBrown did his thing in the **lakers** game !! got ta love him	**neutral**	positive	positive	**neutral**

Table 7.6: Example of output predictions with translated input for targeted cross-lingual sentiment model trained on Arabic and evaluated on identifying sentiment towards targets in English. The target is indicated in bold.

In general, we can see that in both English-to-Arabic and Arabic-to-English transfer, integrating BSW+Lex into the attention model helps it identify more sentiment correctly towards the target. In the third English-to-Arabic example for instance, BSW+Lex is the only model that correctly identifies negative sentiment towards the target ‘occupation’, while both the untargeted and vanilla targeted model predict neutral, failing to recognize ‘occupation’ as a target of sentiment. Similarly, in the first English-to-Arabic example, it is the only model that correctly identifies negative sentiment towards the target ‘george bush’, while the untargeted model predicts positive sentiment, and the vanilla attention model predicts neutral sentiment. Moreover, unlike the other two models, BSW+Lex correctly identifies that the sentiment towards ‘United Nations’ in the first example in Table 7.5 is actually neutral. On the other hand, it misclassifies sentiment towards ‘the United Nations’ in the last example, a more difficult example where ‘United States embassador’, which is the target that receives the negative sentiment, is closely linked to ‘the United Nations’.

We note also that the untargeted model sometimes misses sentiment clues altogether - which we also observed in Chapter 5 when we observed that the untargeted cross-lingual model transferred from Arabic to English makes several errors of the type ‘sentiment indicator missed’. However, the attention mechanism appears to help even in the prediction of untargeted sentiment - such as the last three examples in Table 7.6, where the attention model correctly predicts the *untargeted* sentiment as positive but incorrectly predicts the sentiment towards the target, such as ‘lakers’. In this case, it would seem that the correct neutral prediction made by the untargeted model results from missing the sentiment indicator ‘got ta love him’ rather than attention to the target ‘lakers’. Similarly in the ‘United Nations’ example in Table 7.6, the untargeted model probably misses the sentiment indicator ‘heated attack’.

7.6 Conclusion

This chapter concluded the dissertation’s presentations of models, strategies, and experimental analyses towards identifying sentiment cross-lingually and in low-resource languages. It built on the models, bilingual feature representations, and language-specific considerations presented throughout the work in order to present and evaluate a holistic approach for identifying sentiment towards targets cross-lingually in a low-resource language. Through experiments on English-to-Arabic and Arabic-to-English transfer of targeted sentiment, we were able to demonstrate that our conclusions as to the nature of the bilingual translation corpus, the effect of bilingual sentiment feature representations and target language lexicalization, as well as the selection of morphological pre-processing schemes, hold true for the transfer of targeted sentiment as well as untargeted sentiment.

Furthermore, our results and analysis showed that while a cross-lingual model that attends to sentiment and target states in the sequence can result in some improvements to targeted (as well as untargeted) cross-lingual sentiment prediction, by integrating our bilingual representation features, including bilingual sentiment embeddings and weights updated during training through lexicalization, the performance of the targeted attention model can be substantially improved.

The transfer of targeted sentiment cross-lingually is a very new area of study, and there are several extensions, as well as resources to be created, that can be considered in the future. First, while our choice of language pairs was restricted by the availability of resources - and we have created our Arabic targeted sentiment datasets for this purpose - creating sentiment datasets with target annotations in a number of moderately resourced languages would help further research in this direction by allowing researchers to run targeted sentiment transfer experiments using a larger number of language pairs. Second, targeted cross-lingual sentiment models can be further developed to identify sentiment in open-domain and longer texts, such as those considered in Chapter 6, where there are more candidate targets than what typically exists in tweets. This would enable us to get an even clearer distinction between models that successfully identify sentiment clues versus models that are able to distinguish between sentiment expressed towards different targets. Finally, because of the role that syntax plays in linking target entities with sentiment clues, models that additionally place an emphasis on transfer of syntax, may prove helpful.

Chapter 8

Conclusion

> **"It is good to have an end to journey toward, but it is the journey that matters in the end."**
>
> — Ursula K. Le Guin, *The Left Hand of Darkness*

The ability to identify sentiment in a language with minimal resources is necessary if we are to build natural language processing systems that can aggregate, report, and respond to sentiments expressed in a wide range of genres, scenarios and applications, which include sentiments expressed in customer review texts written in high resource languages, but also sentiments expressed towards entities, issues, and real-life events in regions around the world where hundreds if not thousands of low-resource languages are spoken and written.

This thesis presents resources, techniques, strategies and extensive experimental analyses towards the goal of identifying untargeted and targeted sentiment using cross-lingual means with only labeled data from a high-resource or moderately resourced source language. In contrast to previously published work in the area, our work covered much larger ground in the problem of cross-lingual sentiment analysis; it integrated and demonstrated the effectiveness of untraditional resources such as in-domain comparable corpora and smaller sizes of in-domain parallel corpora as a medium of sentiment transfer, covered 18 target languages from 5 broad language families as well as 15 source languages with a study of the impact of the source language, assessed language-specific morphological representation schemes and integrated support for the identification of sentiment towards a broad range of targets, with cross-lingual targeted experiments demonstrated on two language pairs.

8.1 Contributions

Our work makes several contributions and accompanying findings to the field of cross-lingual sentiment analysis, which we reiterate below.

- **Transfer Model**. We presented and evaluated a cross-lingual model, trained on a high-resource source language and applied directly to a low-resource target language. The model performs effectively, in most cases outperforming baselines and state-of-the-art using only pre-trained cross-lingual word embeddings, but may be further enhanced by lexicalization of the training data into the target language.
- **Bilingual Resources**. We demonstrated the effectiveness of our model using bilingual feature representations from a number of resources, including comparable corpora which have not been used previously for the task of cross-lingual sentiment analysis. These comparable feature representations outperform several baselines as well as, on average, unsupervised methods that rely on the projection of vector spaces of monolingual corpora, and for some languages, even parallel corpora. We showed, additionally, that bilingual feature representations trained on an in-domain parallel corpus result in the best sentiment transfer configuration for most target languages, generally outperforming monolingual-based embedding representations or embedding representations from larger-sized out-of-domain corpora, but that out-of-domain and monolingual corpora are still a viable alternative when in-domain resources are not available.
- **Bilingual Feature Representations**. We presented and evaluated bilingual sentiment embeddings and sentiment output weights which can be used to cre-

ate bilingual sentiment scores, that are pre-trained in a bilingual context on an appropriate translation corpus and integrated in the cross-lingual model. The bilingually trained features outperform sentiment scores which are directly projected from the source language and allow the cross-lingual model to detect more instances of positive and negative sentiment in the target language, especially when the source training data contains label bias. The sentiment embeddings may be updated during training together with target lexicalization, and this configuration results in our best performing model.

- **Source Language**. We studied the impact of the source language on the performance of the cross-lingual model and proposed a method for making use of machine translation amongst high-resource languages to create a parallel corpus between source and target languages in similar language families. Through experiments transferring sentiment from European and Indo-European source languages as well as Arabic and Chinese, we found that the performance of the cross-lingual model is impacted by language family similarity, morphological complexity of the source language vocabulary compared to the target language, and the training set of the source language.
- **Open-domain Targets**. We built rich linguistic models for identifying target entities, not restricted to named entities, along with their sentiment in Arabic open-domain text, and identified the morphological representations and segmentations that work best for identifying targets and sentiment in Arabic. As part of our exploration into the identification of targeted sentiment in open-domain text, we also introduced a new problem: the identification of sentiment towards situation frames, and proposed effective baselines for identifying sentiment towards needs and issues in English and Spanish.
- **Transfer of Targeted Sentiment**. We adapted our cross-lingual model for the transfer of sentiment expressed towards targets, and demonstrated that

our conclusions for cross-lingual transfer of untargeted sentiment hold true for the transfer of targeted sentiment. These include the impact of the in-domain parallel corpus, the effectiveness of target language lexicalization with bilingual sentiment embedding weights and update, and the choice of morphological segmentation schemes.

- **Resource Contributions**. Our work makes a number of resource contributions, which include a comparable corpus of topic-aligned and article-aligned Wikipedia articles for 17 languages, three sentiment datasets for Arabic untargeted and targeted sentiment for the genres of Twitter and online newspaper comments, the latter crowdsourced using Amazon Mechanical Turk and providing sentiment annotations for over 4000 target entities. Finally, we make three native-annotated sentiment evaluation datasets available for Uyghur, Tigrinya, and Sinhalese.

8.2 Scope

As was mentioned at several instances in the thesis chapters, our work focused on the classification of sentiment into three classes: positive, negative, and neutral, with additional two-class experiments included when evaluating the identification of targeted sentiment. Our work does not distinguish between neutral expressions of sentiment and non-polar expressions of opinion or subjectivity, such as expressions of judgment, belief, or surprise. Differentiating between expressions of subjectivity has been studied in the past for the English language, e.g by Wiebe et al. (2005), and is out of the scope of our cross-lingual work. However, our cross-lingual sentiment models, bilingual sentiment embeddings, and extensive feature evaluations provide a basis for developing cross-lingual models of opinions and subjectivity with finer-grained categories.

Similarly, our work on targeted sentiment focused on targets of sentiment, but because the genres of text tackled in the thesis focused on social media, customer reviews, and forums for online discussion or commentary, we have assumed that in the majority of cases the source of the expressed sentiment is the author of the text. However, in other genres, such as news articles, this may not be the case, and the development of cross-lingual models for the identification of sentiment sources will be a helpful future direction for all genres of text. Our Arabic open domain dataset, for instance, may be extended for the annotation of source entities for all targets where the source is not the author.

Our cross-lingual sentiment models are able to operate under a number of resources: parallel, comparable, or monolingual; only the pre-trained cross-lingual embeddings need to be provided. However, one limitation of our bilingual sentiment embeddings is that because they rely on translation context, they require a parallel or comparable corpus; they cannot, for example, be used in conjunction with purely monolingual methods such as VECMAP. To do this, it would require developing a version of our bilingual sentiment embedding approach and of VECMAP that incorporates sentiment information into the mapping of similarity matrices across languages. Similarly, our comparable corpus and the embeddings created from it were not applicable to Tigrinya, because its Wikipedia corpus contained very few articles aligned with English. Instead, we relied on using a very small parallel corpus, which did well for this language.

8.3 Future Directions and Applications

Through the work set forth in the thesis, we aim to encourage future studies in the fields of cross-lingual sentiment analysis, cross-lingual targeted sentiment analysis, and related applications that can make use of our models and resources.

With regard to untargeted cross-lingual sentiment analysis, having demonstrated the viability of using comparable corpora as a medium of sentiment transfer, future work could involve the collection of comparable translation corpora from Twitter or other social media sites, with the goal of capturing bilingual representations of context and sentiment that occur in informal, mispelled, or dialectal text, which in our analysis was observed to be a source of error due to resulting out-of-vocabulary words. While Twitter embeddings and Twitter sentiment embeddings exist monolingually, they have not been developed bilingually and a corpus could be created by searching social media websites bilingually for keywords and hashtags related to specific topics or events in a similar manner to which our Wikipedia corpus was created. The development of this corpus would require creative techniques - cross-lingual topic models, or multi-word translations using bilingual dictionaries for example - to identify comparable tweets in the source and target languages, as well as a method for keyword tweet search in languages not supported by the Twitter search API (Tigrinya, for example, is one such language).

Additionally, recent developments in bidirectional language-modeling-based techniques such as BERT (Devlin et al., 2018) have led to significant successes for several downstream natural language processing tasks, such as question answering, natural language inference, and textual entailment. Using BERT pre-training and fine-tuning in our sentiment transfer models is another avenue for future work. BERT would be best coupled with code-switched and mixed-language-document text representations (e.g the dictionary-code-switched embeddings and comparable embeddings derived using merged shuffling). However, it bears mentioning that BERT requires substantially large pre-training corpora (in the order of 2000M words), which is not likely to be available for our low-resource languages.

With regard to targeted cross-lingual sentiment analysis, future work involves running cross-lingual targeted experiments on a large set of language pairs, and would likely require the use of native informants for the annotation of targeted evaluation datasets for a number of low-resource languages. A valuable future direction in targeted cross-lingual sentiment analysis would be the development of models that identify sentiment towards situations in a low-resource language. Such models could make use of our cross-lingual targeted attention model, for example, to identify sentiment expressed towards situation frame keywords. Another direction would be the development of a joint model that identifies situation frames and sentiment expressed towards them. As with all the tasks considered in our work, the development of such a model requires the creation of resources - namely, the annotation of high-resource training data for the identification of situation frames and their sentiment, and a small amount of evaluation data annotated for situation frames and their sentiment in the target language.

Our work in Arabic and other moderately-resourced languages - in particular, the successful transfer of sentiment amongst the Slavic languages, Germanic languages, and between Arabic and Tigrinya, showed the promise of using moderately-resourced languages as an alternative to English when transferring sentiment to low-resource target languages in the same language family. However, this also requires the dedication of efforts to create resources, namely training data, for moderately-resourced languages. We believe such a direction is feasible because it would involve less challenges compared with annotating training datasets for low-resource languages, where a fewer number of native speakers are accessible on crowdsourcing platforms, for example.

On a final note, we mention a number of other applications that would benefit from our models, feature representations, and resources, in the development of cross-lingual natural language processing systems. One is the cross-lingual detection of emotion, a task that would also be valuable when assessing the responses of populations to natural disasters and political incidents. Our cross-lingual sentiment models have been successfully used in conjunction with the emotion detection system of Tafreshi and Diab (2018) to help identify situations in text that could contain emotions such as joy, fear, or anger. Another application is the cross-lingual detection of urgency in such situations, where our model predictions have been used as a feature integrated in an urgency detection system. Finally, we hope our bilingual sentiment embeddings and targeted sentiment models would be used in applications that ensure that social media is a safe and courteous place to conduct discussions; these may include, for example, the identification of offensive language in social media directed at individuals or groups (Zampieri et al., 2019a,b), or the cross-lingual transfer of such models for the purpose of identifying hate speech expressed in different languages.

Bibliography

Muhammad Abdul-Mageed and Mona T. Diab. Subjectivity and sentiment annotation of modern standard Arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 110–118. Association for Computational Linguistics, 2011.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28 (1):20–37, 2014.

Amjad Abu-Jbara, Ben King, Mona T Diab, and Dragomir R Radev. Identifying opinion subgroups in Arabic online discussions. In *ACL (2)*, pages 829–835, 2013.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

Md S Akhtar, Palaash Sawant, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. Association for Computational Linguistics, 2018.

Mohammad Al-Smadi, Omar Qawasmeh, Bashar Talafha, and Muhannad Quwaider. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 726–730. IEEE, 2015.

Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of computational science*, 27:386–393, 2018.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*, 2018.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. A large scale Arabic sentiment lexicon for Arabic opinion mining. *ANLP 2014*, pages 165–173, 2014.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Alexandra Balahur and Marco Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75, 2014.

Jeremy Barnes, Patrik Lambert, and Toni Badia. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, 2016.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Bilingual sentiment embeddings: Joint projection of sentiment across languages. *arXiv preprint arXiv:1805.09016*, 2018.

Peter Baumann and Janet B Pierrehumbert. Using resource-rich languages to improve morphological analysis of under-resourced languages. In *LREC*, pages 3355–3359, 2014.

Yassine Benajiba, Mona Diab, and Paolo Rosso. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 284–293. Association for Computational Linguistics, 2008.

Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. Multivec: a multilingual and multilevel representation learning toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC)*, 2016.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics, 2012.

Prakhar Biyani, Cornelia Caragea, and Narayan Bhamidipati. Entity-specific sentiment classification of Yahoo news comments. *arXiv preprint arXiv:1506.03775*, 2015.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. Introducing the Arabic Wordnet project. In *Proceedings of the third international WordNet conference*, pages 295–300. Citeseer, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics, 2010.

John D Byrum. ISO 639-1 and ISO 639-2: International standards for language codes. ISO 15924: International standard for names of scripts. 1999.

Paula Carvalho, Luís Sarmento, Jorge Teixeira, and Mário J Silva. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 564–568. Association for Computational Linguistics, 2011.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-09-1. URL `http://dl.acm.org/citation.cfm?id=1626394.1626430`.

Aditi Chaudhary, Siddharth Dalmia, Junjie Hu, Xinjian Li, Austin Matthews, Aldrian Obaja Muis, Naoki Otani, Shruti Rijhwani, Zaid Sheikh, Nidhi Vyas, et al. The ARIEL-CMU systems for LoReHLT18. *arXiv preprint arXiv:1902.08899*, 2019.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*, 2016.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018.

Caitlin Christianson, Jason Duncan, and Boyan Onyshkevych. Overview of the DARPA LORELEI program. *Machine Translation*, 32(1-2):3–9, 2018.

Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, pages 1–21, 2014.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, Jennifer Tracey, K Choukri, T Declerck, S Goggi, M Grobelnik, et al. Selection criteria for low resource language programs. In *LREC*, 2016.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

Lingjia Deng and Janyce Wiebe. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 179–189, 2015a.

Lingjia Deng and Janyce Wiebe. MPQA 3.0: An entity/event-level sentiment corpus. 2015b.

Michael Denkowski, Hassan Al-Haj, and Alon Lavie. Turker-assisted paraphrasing for English-Arabic machine translation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 66–70. Association for Computational Linguistics, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 49–54, 2014.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 429–433, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6.

Ayman El-Kilany, Amr Azzam, and Samhaa R El-Beltagy. Using deep neural networks for extracting sentiment targets in arabic tweets. In *Intelligent Natural Language Processing: Trends and Applications*, pages 3–15. Springer, 2018.

Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. A machine learning approach for opinion holder extraction in arabic language. *arXiv preprint arXiv:1206.1011*, 2012.

Noura Farra and Kathleen McKeown. SMARTies: Sentiment models for Arabic target entities. In *Proceedings of the European Chapter of the Association for Computational Lingustics (EACL*, 2017.

Noura Farra and Kathleen McKeown. Practical pre-trained embeddings for cross-lingual sentiment analysis. In *Submission to the 57th Annual Meeting of the Association of Computational Linguistics*, 2019.

Noura Farra, Nadi Tomeh, Alla Rozovskaya, and Nizar Habash. Generalized character-level spelling error correction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 161–167, 2014.

Noura Farra, Kathleen McKeown, and Nizar Habash. Annotating targets of opinions in Arabic using crowdsourcing. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 89–98, 2015a.

Noura Farra, Swapna Somasundaran, and Jill Burstein. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74, 2015b.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E14-1049`.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Spence Green and Christopher D Manning. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 394–402. Association for Computational Linguistics, 2010.

Nizar Habash. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187, 2010.

Nizar Habash and Fatiha Sadat. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52. Association for Computational Linguistics, 2006.

Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

Chiara Higgins, Elizabeth McGrath, and Lailla Moretto. MTurk crowdsourcing: a viable method for rapid discovery of Arabic nicknames? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 89–92. Association for Computational Linguistics, 2010.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Christopher Hogan. OCR for minority languages. In *Symposium on Document Image Understanding Technology*, 1999.

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. SentiPers: A sentiment analysis corpus for Persian, 2015.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Ozan Irsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In *EMNLP*, pages 720–728, 2014.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.

Joo Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler Lussier. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, 2017.

Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics, 2006.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Roman Klinger and Philipp Cimiano. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, 2015.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Taku Kudo. Crf++: Yet another CRF toolkit. *Software available at http://crfpp. sourceforge. net*, 2005.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79. Association for Computational Linguistics, 2010.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436, 2015.

Young-Suk Lee. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60. Association for Computational Linguistics, 2004.

M Paul Lewis. *Ethnologue: Languages of the world*. SIL international, 2009.

Bo Li and Eric Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International conference on computational linguistics*, pages 644–652. Association for Computational Linguistics, 2010.

Yiou Lin, Hang Lei, Jia Wu, and Xiaoyu Li. An empirical study on sentiment classification of Chinese review using word embedding. *arXiv preprint arXiv:1511.01665*, 2015.

Janna Lipenkova. A system for fine-grained aspect-based sentiment analysis of Chinese. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 55–60, 2015.

Patrick Littell, Tian Tian, Ruochen Xu, Zaid Sheikh, David Mortensen, Lori Levin, Francis Tyers, Hiroaki Hayashi, Graham Horwood, Steve Sloto, et al. The ARIEL-CMU situation frame detection pipeline for LoReHLT16: a model translation approach. *Machine Translation*, 32(1-2):105–126, 2018.

Jiangming Liu and Yue Zhang. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 572–577, 2017.

Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 1433–1443, 2015.

Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

Mike Maxwell and Baden Hughes. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the workshop on frontiers in linguistically annotated corpora 2006*, pages 29–37. Association for Computational Linguistics, 2006.

Andrew K McCallum. {MALLET: A Machine Learning for Language Toolkit}. 2002.

Karine Megerdoomian and Dan Parvaz. Low-density language bootstrapping: the case of Tajiki Persian. In *LREC*, 2008.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–581, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

George A Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Margaret Mitchell, Jacqueline Aguilar, Theresa Wilson, and Benjamin Van Durme. Open domain targeted sentiment. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, 2013.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, volume 16, pages 31–41, 2016a.

Saif Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130, 2016b.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar, October 2014. Association for Computational Linguistics.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. Twitter sentiment for 15 European languages, 2016. Slovenian language resource repository CLARIN.SI.

Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy. Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 70–82, 2018.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June 2016. Association for Computational Linguistics.

Islam Obaidat, Rami Mohawesh, Mahmoud Al-Ayyoub, Mohammad AL-Smadi, and Yaser Jararweh. Enhancing the determination of aspect categories and their polarities in Arabic reviews using lexicon-based approaches. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on*, pages 1–6. IEEE, 2015.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *LREC*, volume 14, pages 1094–1101, 2014.

Haiyun Peng, Yukun Ma, Yang Li, and Erik Cambria. Learning multi-grained aspect target sequence for Chinese sentiment analysis. *Knowledge-Based Systems*, 148: 167–176, 2018.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Maria Pontiki, Haris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, 2014.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30, 2016.

Mohammad Sadegh Rasooli and Michael Collins. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, 2015.

Mohammad Sadegh Rasooli and Michael Collins. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293, 2017. ISSN 2307-387X.

Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, Tao Yu, and Kathleen McKeown. Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1-2):143–165, 2018.

Leanne Rolston and Katrin Kirchhoff. Collection of bilingual data for lexicon transfer learning. Technical report, Technical Report UW-EE-2016-0001, 2016.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. Semeval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 451–463, Denver, Colorado, USA, 2015a.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June 2015b. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. Finding the sources and targets of subjective expressions. In *LREC*, pages 2781–2788, 2008.

Mohammad Salameh, Saif M. Mohammad, and Svetlana Kiritchenko. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, 2015.

Iman Saleh, Alessandro Moschitti, Preslav Nakov, Lluís Màrquez, and Shafiq Joty. Semantic kernels for semantic parsing. In *EMNLP*, pages 436–442, 2014.

Anas Shahrour, Salam Khalifa, and Nizar Habash. Improving Arabic diacritization through syntactic analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1315, 2015.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics, 2009.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Discourse level opinion relations: An annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 129–137. Association for Computational Linguistics, 2008.

Veselin Stoyanov and Claire Cardie. Annotating topics of opinions. In *LREC*, 2008.

Karl Stratos, Do-kyum Kim, Michael Collins, and Daniel Hsu. A spectral algorithm for learning class-based n-gram models of natural language. *Proceedings of the Association for Uncertainty in Artificial Intelligence*, 2014.

Afraz Z Syed, Muhammad Aslam, and Ana Maria Martinez-Enriquez. Associating targets with sentiunits: a step forward in sentiment analysis of urdu text. *Artificial intelligence review*, 41(4):535–561, 2014.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 477–487. Association for Computational Linguistics, 2012.

Shabnam Tafreshi and Mona Diab. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913, 2018.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565, 2014.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*, 2015.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509, 2016.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584. Association for Computational Linguistics, 2010.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*, 2016.

Ivan Vulić and Marie-Francine Moens. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994, 2016.

Xiaojun Wan. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2009)*, pages 235–243, Singapore, SN, 2009.

Yequan Wang, Minlie Huang, Li Zhao, and Xiaoyun Zhu. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

Theresa Ann Wilson. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states.* ProQuest, 2008.

Bishan Yang and Claire Cardie. Joint inference for fine-grained opinion extraction. In *ACL (1)*, pages 1640–1649, 2013.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics, 2000.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, 2017.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. Large scale Arabic error annotation: Guidelines and framework. In *LREC*, pages 2362–2369, 2014.

Omar F Zaidan and Chris Callison-Burch. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics, 2011.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, 2019a.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019b.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59. Association for Computational Linguistics, 2012.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on EMNLP*, pages 612–621, 2015.

Lin Zheng, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1089–1098. ACM, 2014.

Guangyou Zhou, Tingting He, and Jun Zhao. Bridging the language gap: Learning distributed semantics for cross-lingual sentiment classification. In Chengqing Zong, Jian-Yun Nie, Dongyan Zhao, and Yansong Feng, editors, *Natural Language Processing and Chinese Computing*, pages 138–149, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

HuiWei Zhou, Long Chen, Fulin Shi, and Degen Huang. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 430–440, Beijing, China, July 2015. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Attention-based LSTM network for cross-lingual sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256, Austin, Texas, November 2016. Association for Computational Linguistics.

Ayah Zirikly and Masato Hagiwara. Cross-lingual transfer of named entity recognizers without parallel corpora. *Volume 2: Short Papers*, pages 390–396, 2015.