SMARTies: Sentiment Models for Arabic Target entities

Noura Farra and Kathleen McKeown
Columbia University
Department of Computer Science
{noura,kathy}@cs.columbia.edu

Abstract
We consider entity-level sentiment analysis in Arabic, a morphologically rich language with increasing resources. We present a system that is applied to complex posts written in response to Arabic newspaper articles. Our goal is to identify important entity “targets” within the post along with the polarity expressed about each target. We achieve significant improvements over multiple baselines, demonstrating that the use of specific morphological representations improves the performance of identifying both important targets and their sentiment, and that the use of distributional semantic clusters further boosts performances for these representations, especially when richer linguistic resources are not available.

1 Introduction
Target-specific sentiment analysis has recently become a popular problem in natural language processing. In interpreting social media posts, analysis needs to include more than just whether people feel positively or negatively; it also needs to include what they like or dislike. The task of finding all targets within the data has been called “open-domain targeted sentiment” (Mitchell et al., 2013; Zhang et al., 2015). If we could successfully identify the targets of sentiment, it would be valuable for a number of applications including sentiment summarization, question answering, understanding public opinion during political conflict, or assessing needs of populations during natural disasters.

In this paper, we address the open-domain targeted sentiment task. Input to our system consists of online posts, which can be comprised of one or multiple sentences, contain multiple entities with different sentiment, and have different domains. Our goal is to identify the important entities towards which opinions are expressed in the post; these can include any nominal or noun phrase, including events, or concepts, and they are not restricted to named entities as has been the case in some previous work. The only constraint is that the entities need to be explicitly mentioned in the text. Our work also differs from much work on targeted sentiment analysis in that posts are long, complex, with many annotated targets and a lack of punctuation that is characteristic of Arabic online language. Figure 1 shows an example post, where targets are either labeled positive (green) if a positive opinion is expressed about them and negative (yellow) if a negative opinion is expressed.

To identify targets and sentiment, we develop two sequence labeling models, a target-specific model and a sentiment-specific model. Our models try to learn syntactic relations between entities and opinion words, but they also make use of (1) Arabic morphology and (2) entity semantics. Our use of morphology allows us to capture all “words” that play a role in identification of the target, while our use of entity semantics allows us to group together similar entities which may all be targets of the same sentiment; for example, if a commenter expresses negative sentiment towards...
the United States, they may also express negative sentiment towards America or Obama.

Our results show that morphology matters when identifying entity targets and the sentiment expressed towards them. We find for instance that the attaching Arabic definite article Al+ 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. We also conduct a detailed analysis of errors revealing that the task generally entails hard problems such as a considerable amount of implicit sentiment and the presence of multiple targets with varying importance.

In what follows, we describe related work (§ 2), data and models (§ 3 and § 4), and linguistic decisions made for Arabic (§ 5). In § 6, we describe our use of word vector clusters learned on a large Arabic corpus. Finally, § 7 presents experiments and detailed error analysis.

2 Related Work

Aspect-based and Entity-specific Analysis

Early work in target-based sentiment looked at identifying aspects in a restricted domain: product or customer reviews. Many of these systems used unsupervised and topical methods for determining aspects of products; Hu and Liu (2004) used frequent feature mining to find noun phrase aspects, 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.

Entity-specific sentiment analysis has been frequently studied in 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. Biyani et al. (2015) studied sentiment towards entities in online posts, where 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. More recent work uses LSTM and RNN networks to determine sentiment toward aspects in product reviews (Wang et al., 2016) and 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., 2016) towards pre-defined topics in Twitter, both on English data.

Open domain targeted analysis

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.") Recent work (Deng and Wiebe, 2015) identifies entity sources and targets, as well as the sentiment expressed by and towards these entities. This work was based on probabilistic soft logic models, also with a focus on direct subjective expressions.

There is also complementary work on using neural networks for tagging open-domain targets (Zhang et al., 2015; Liu et al., 2015) in shorter posts. Previous work listed did not consider word morphology, or explicitly model distributional entity semantics as indicative of the presence of sentiment targets.

Related work in Arabic

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 as we do in our sequence labeling task. Analysis of such segmentation schemes has not been reported for Arabic sentiment tasks, which cover mostly sentence-level sentiment analysis and where the lemma or surface bag-of-word representations have typically been sufficient.

There are now many studies on sentence-level
sentiment analysis in Arabic news and social media (Abdul-Mageed and Diab, 2011; Mourad and Darwish, 2013; Refaee and Rieser, 2014; Salameh et al., 2015). Elarnaoty et al. (2012) proposed identifying sources of opinions in Arabic using a CRF with a number of patterns, lexical and subjectivity clues; 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. There are no other sentiment target studies in Arabic that we know of. In our experiments, we compare to methods similar to these baseline systems, as well as to results of English work that is comparable to ours.

Entity Clusters 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); we are not aware of such work for identifying entity targets of sentiment.

3 Data
We use the Arabic Opinion Target dataset developed by Farra et al. (2015), which is publicly available\(^1\). The data consists of 1177 online comments posted in response to Aljazeera Arabic newspaper articles and is part of the Qatar Arabic Language Bank (QALB) corpus (Habash et al., 2013; Zaghouani et al., 2014). The comments are 1-3 sentences long with an average length of 51 words. They were selected such that they included topics from three domains: politics, culture, and sports.

Targets are always noun phrases and they are either labeled positive if a positive opinion is expressed about them and negative if a negative opinion is expressed (as shown in Figure 1). Targets were identified using an incremental process where first important entities were identified, and then entities agreed to be neutral were discarded (the annotation does not distinguish between neutral and subjective neutral).

The data also contains ambiguous or ‘undeter-

\(^1\)www.cs.columbia.edu/~noura/Resources.html

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

Table 1: Example of CRF annotations.

mined’ targets where annotators did agree they were targets, but did not agree on the polarity. We use these targets for training our target model, but discard them when training our sentiment polarity model. There are 4886 targets distributed as follows: 38.2% positive, 50.5% negative, and 11.3% ambiguous. We divide the dataset into a training set (80%), development set (10%), and blind test set (10%), all of which represent the three different domains. We make the splits available for researchers to run comparative experiments.

4 Sequence Labeling Models
For modeling the data, we choose Conditional Random Fields (CRF) (Lafferty et al., 2001) for the ability to engineer Arabic linguistic features and because of the success of CRF models in the past for entity identification and classification related tasks.

We build two linear chain CRF models:

1. **Target Model** This model predicts a sequence of labels \(\vec{E}\) for a sequence of input tokens \(\vec{x}\), where

\[ E_i \in \{ T(\text{target}), O(\text{not\_target}) \} \]

and each token \(x_i\) is represented by a feature vector \(\vec{f}_{it}\). A token is labeled \(T\) 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 \(\vec{S}\) for the sequence \(\vec{x}\),

\[ S_i \in \{ P(\text{pos}), N(\text{neg}), \emptyset(\text{neutral}) \} \]

and each token \(x_i\) is represented by a feature vector:

\[ (\vec{f}_{is}, E_i); E_i \in \{ T, O \} \]

Additionally, this model has the constraint:

\[ \text{if } E_i = T, S_i \in \{ P, N \} \]

and otherwise

\[ S_i = \emptyset \]
The last constraint indicating that sentiment is either positive or negative is ensured by the training data, where we have no examples of target tokens having neutral sentiment. The two models are trained independently. Thus, if target words 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 whereby the entire sequence is optimized to predict sentiment polarities for the targets. An example sequence is shown in Table 1, where the dictator is an entity target towards which the writer implicitly expresses negative sentiment.

5 Arabic Morphology and Linguistics

5.1 Arabic Morphology

In Arabic, clitics and affixes can attach to the beginning and end of the word stem, making words complex. For example, in the sentence ‘So they welcomed her’, the discourse conjunction (so \(\text{ض} \)), the opinion target (her \(\text{هي} \)), opinion holder (they \(\text{هم} \)), and the opinion expression itself (welcomed \(\text{استقبلوها} \)) are all collapsed in the same word.

Clitics, such as conjunctions +و \(w^+\), prepositions +ب \(b^+\), the definite article \(\text{اليال} Al^+ \) ‘the’ (all of which attach at the beginning), and possessive pronouns and object pronouns +ه \(h^+\), +ها \(hA^+\) ‘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.

The morphological analyzer MADAMIRA (Pasha et al., 2014) enables the tokenization of a word using multiple schemes. We consider the following two schemes:

- **D3**: the Declitization scheme which splits off conjunction clitics, particles and prepositions, \(\text{AI}^+\), and all the enclitics at the end.
- **ATB**: the Penn Arabic Treebank tokenization, which separates all clitics above except the definite article \(\text{AI}^+\), which it keeps attached.

For a detailed description of Arabic concatenative morphology and tokenization schemes, the reader is referred to Habash (2010). For each token, we add a part of speech feature. For word form (non-clitic) tokens, we use the part of speech (POS) feature produced by the morphological analyzer. We consider the surface word and the lemma for representing the word form. For the clitics that were split off, we use a detailed POS feature that is also extracted from the output of the analyzer and can take such forms as \(\text{DET} \) for \(\text{AI}^+\) or \(\text{poss} \_\text{pron} \_\text{3MP} \) for third person masculine possessive pronouns. Table 2 shows the words and part of speech for the input sentence فاستقبلوها ‘so they welcomed her’ فاستقبلوها, using the lemma representation for the word form and the D3 tokenization scheme.

These lexical and POS features are added to both our target model and sentiment model.

5.2 Sentiment Features

The choice of sentiment lexicon is an important consideration when developing systems for new and/or low-resource languages. We consider three lexicons: (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 and an Arabic lexical database (Badaro et al., 2014), and (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.

For the target model, we add token-level binary features representing subjectivity, and for the sentiment model, we add both subjectivity and polarity features.

We also add 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.

5.3 Syntactic Dependencies

We ran the CATiB (Columbia Arabic Treebank) dependency parser (Shahrour et al., 2015) on our data. CATiB uses a number of intuitive labels specifying the token’s syntactic role: e.g. \(\text{SBJ}, \text{OBJ}, \text{MOD}, \) and \(\text{IDF} \) for the Arabic idafa construct (e.g. رئيس الحكومة ‘president of government’), as well as its part of speech role. In addition to the sentiment dependency features specifying the sentiment of parent words, we added 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_vrb or nom_idf_nom, as well as the sentiment path from the token to the parent, e.g. nom(neutral)_obj_vrb(negative).

5.4 Chunking and Named Entities

The morphological analyzer MADAMIRA also produces base phrase chunks (BPC) and named entity tags (NER) for each token. We add features for these as well, based on the hypothesis that they will help define the spans for entity targets, whether they are named entities or any noun phrases. We refer to the sentiment and target models that utilize Arabic morphology, sentiment, syntactic relations and entity chunks as best-linguistic.

6 Word Clusters and Entity Semantics

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. They may repeat frequently in a post 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, Obama, the Americans, Westerners}. Such entities can cluster together semantically and it is likely that a person 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 with a reduced feature space compared to Arabic lexical features. 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 word embedding vectors using the skip-gram method (Mikolov et al., 2013) and cluster them using the K-Means algorithm (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.

The vectors are built on Arabic Wikipedia on a corpus of 137M words resulting in a vocabulary of 254K words. We preprocess the corpus by tokenizing (using the schemes described in section 5) and lemmatizing before building the word vectors. We vary the number of clusters and use the clusters as binary features in our target and sentiment models.

7 Experiments and Results

7.1 Experiments

Setup

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

We evaluate all our experiments on the development set which contains 116 posts and 442 targets, and present a final result with the best models on the unseen test. For the SentiWordNet-based lexicon ArSenL, we tune for the sentiment score threshold and use $t=0.2$. We use Google’s word2vec tool for building and clustering word vectors with dimension 200. We vary the number of clusters $k$ between 10 (25K words/cluster) and 20K (12 words/cluster).

Baselines

For evaluating the predicted targets, we follow work in English (Deng and Wiebe, 2015) and use the all-NP baseline, where all nouns and noun phrases in the post are predicted as important targets.

For evaluating sentiment towards targets, we consider four baselines: the majority baseline which always predicts negative, and the lexicon

---

2https://dumps.wikimedia.org/arwiki/20160920/arwiki-20160920-pages-articles.xml.bz2
3https://github.com/dav/word2vec
baseline evaluated in the case of each of our three lexicons: manually created, WordNet-based, and English-translated. The strong lexicon baseline splits the post 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).

We run our pipelined models for all morphological representation schemes: surface word (no token splits), lemma (no clitics), lemma with ATB clitics (contain all token splits except Al+), and lemma with D3 clitics (contains all token splits). We explore the effect of semantic word clusters in these scenarios. Finally we show our best-linguistic (high-resource) model, and the resulting integration with word clusters.

7.2 Results

Tables 3-5 show the results. Target F-measure is calculated using the subset metric (similar to metrics 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. Egyptian’s position and Israel’s position do not match). For this task, in the case of multiple mentions of the same entity in the post, any mention will be considered correct if the subset matches\(^4\) (e.g. if فلسطين is a gold target, and دولة فلسطين is predicted at a different position in the post, it is still correct). This evaluation is driven from the sentiment summarization perspective: we want to predict the overall opinion in the post towards an entity.

\(F\text{-pos}, F\text{-neg}, \text{and Acc-sent}\) show the performance of the sentiment model on only the correctly predicted targets\(^5\). Since the target and sentiment models are trained separately, this is meant to give an idea of how the sentiment model would perform in standalone mode, if targets were already provided.

\(F\text{-all}\) shows the overall F-measure showing the performance of correctly predicted targets with correct sentiment compared to the total number of polar targets. This evaluates the end-to-end scenario of both important target and sentiment prediction.

Best results are shown in bold. Significance thresholds are calculated for the best performing systems (Tables 4-5) using the approximate randomization test (Yeh, 2000) for target recall, precision, F-measure, Acc-sent and F-all. 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 \(p < 0.05\). Our dataset is small; nonetheless we get significant results.

Comparing Sentiment Lexicons Table 3 shows the results comparing the different baselines. All targets are retrieved using all-NP; sentiment is determined using the lexical baselines. As expected, the all-NP baseline shows near perfect recall and low precision in predicting important targets. We observe that the gloss-translated MPQA lexicon outperforms the two other Arabic lexicons among the sentiment baselines.

We believe that the hit rate of MPQA is higher than that of the smaller, manually-labeled SIFAAT, and it is more precise than the automatically generated WordNet-based lexicon ArSenL. The performance of MPQA is, however, reliant on the availability of high-quality English glosses. We found MPQA to consistently outperform in the model results, so in our best-linguistic models, we only show results using the MPQA lexicon.

Comparing Morphology Representations Looking at table 4, we can see that using the lemma representation easily outperforms the sparser surface word, and that adding tokenized clitics as separate tokens outperforms representations which only use the word form. Moreover, upon using the D3 decliticization method, we observe a significant increase in recall of targets over the ATB representation. This shows that the presence of the Arabic definite article آل Al+ is an important indicator of a target entity; thus, even if an entity is not named, Al+ indicates that it is a known entity and is likely more salient.

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 score and the D3 representation has the lowest af-

\(^4\)We have also computed the performance for mention-overlap; the difference in target F-measure is 2 points and consistent across the different systems.

\(^5\)We exclude targets with ambiguous sentiment whose polarity was not agreed on by the annotators.
Table 3: Target and sentiment results using baselines; all-NP for targets and lexicons for sentiment.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Target Recall</th>
<th>Target Precision</th>
<th>Target F-score</th>
<th>Sentiment F-pos</th>
<th>Sentiment F-neg</th>
<th>Sentiment Acc-sent</th>
<th>Sentiment F-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline1</td>
<td>98.4</td>
<td>29.2</td>
<td>45</td>
<td>0</td>
<td>72.4</td>
<td>56.8</td>
<td>12.4</td>
</tr>
<tr>
<td>Baseline2</td>
<td>98.4</td>
<td>29.2</td>
<td>45</td>
<td>50.6</td>
<td>64.3</td>
<td>58.6</td>
<td>12.7</td>
</tr>
<tr>
<td>Baseline3</td>
<td>98.4</td>
<td>29.2</td>
<td>45</td>
<td>61</td>
<td>58</td>
<td>59.5</td>
<td>13.1</td>
</tr>
<tr>
<td>Baseline4</td>
<td>98.4</td>
<td>29.2</td>
<td>45</td>
<td>67</td>
<td>63.7</td>
<td>65.4</td>
<td>14.2</td>
</tr>
</tbody>
</table>

Table 4: Target and sentiment results using different morphological representations. All models use POS.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Target Recall</th>
<th>Target Precision</th>
<th>Target F-score</th>
<th>Sentiment F-pos</th>
<th>Sentiment F-neg</th>
<th>Sentiment Acc-sent</th>
<th>Sentiment F-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface + POS</td>
<td>41</td>
<td>60.6</td>
<td>48.9</td>
<td>62.2</td>
<td>73.6</td>
<td>68.9</td>
<td>32.6</td>
</tr>
<tr>
<td>Lemma + POS</td>
<td>48.2**</td>
<td>60.5</td>
<td>53.7*</td>
<td>65.4</td>
<td>77.6</td>
<td>72.8</td>
<td>38.1**</td>
</tr>
<tr>
<td>+ATB tokens</td>
<td>52.4*</td>
<td>59.5</td>
<td>55.7</td>
<td>61.3</td>
<td>75.7</td>
<td>70.1</td>
<td>38.2</td>
</tr>
<tr>
<td>+D3 tokens</td>
<td>59.6**</td>
<td>55.7*</td>
<td>57.6</td>
<td>64.1</td>
<td>73</td>
<td>69.2</td>
<td>36.1</td>
</tr>
</tbody>
</table>

Effect of Word Clusters

Figures 2 - 5 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.) 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 4; 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, 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 sentiment is less clear; it seems to benefit the D3 and ATB schemes more than lemma (significant boosts in sentiment accuracy). The improvements in F-measure and F-all observed by using the best value of $k$ is statistically significant for all schemes ($k=10$ for lemma, $k=250$ for lemma+ATB, $k=500$ for lemma+D3, with F-all values of 40.7, 41.5, and 39.1 respectively). 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).

Performance of Best Linguistic Model

Table 5 shows the performance of our best-linguistic model, which in addition to the word form and part of speech, contains named entity and base phrase chunks, the syntactic dependency features, and the sentiment lexicon features. The best linguistic model is run using both ATB and D3 tokenization schemes, and then using a combined ATB+D3 scheme where we use D3 for the target model and remove the extra clitics before piping in the output to the sentiment model. This combined
Table 5: Performance of best linguistic model

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
<th>F-pos</th>
<th>F-neg</th>
<th>Acc-sent</th>
<th>F-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>best-linguistic-ATB</td>
<td>53</td>
<td>62.1</td>
<td>57.2</td>
<td>68.6</td>
<td>79.4</td>
<td>75.1</td>
<td>40.7</td>
</tr>
<tr>
<td>best-linguistic-D3</td>
<td>64.2***</td>
<td>58.8</td>
<td>61.4*</td>
<td>62.7</td>
<td>75.6</td>
<td>70.5***</td>
<td>39.1</td>
</tr>
<tr>
<td>best-linguistic-D3+ATB</td>
<td>63.7</td>
<td>58.8</td>
<td>61.4</td>
<td>67.7</td>
<td>80</td>
<td>75.4***</td>
<td>43.3***</td>
</tr>
<tr>
<td>best-linguistic+clusters</td>
<td>66.2</td>
<td>57.8</td>
<td>61.8</td>
<td>70</td>
<td>80</td>
<td>76</td>
<td>44.2</td>
</tr>
</tbody>
</table>

Table 6: Target and sentiment results on test data.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>Acc</th>
<th>F-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best-D3</td>
<td>63.7</td>
<td>52.3</td>
<td>57.4</td>
<td>69.4</td>
<td>35.4</td>
</tr>
<tr>
<td>Best-D3+ATB</td>
<td>63.7</td>
<td>51.8</td>
<td>57.1</td>
<td>70.3</td>
<td>36.8</td>
</tr>
<tr>
<td>+clusters</td>
<td>65.6</td>
<td>50.2</td>
<td>56.9</td>
<td>73.6</td>
<td>38.1</td>
</tr>
</tbody>
</table>

Figure 4: Target F-score vs clusters.

Figure 5: Sentiment accuracy vs clusters.

Figure 6: Overall F-score vs clusters.

scheme results in the best results overall: F-score of 61.4 for targets, accuracy of 75.4 for sentiment and overall F-measure of 43.1.

Adding the richer linguistic resources results in both improved target precision, recall, and sentiment scores, with F-measure for positive targets reaching 67.7 for positive targets and 80 for negative targets. Performance exceeds that of the simpler models which use only POS and word clusters, but it is worth noting that using only the basic model with the word clusters can achieve significant boosts in recall and F-measure bringing it closer to the rich linguistic model.

The last row shows the best linguistic model D3+ATB combined with the clusters (best result for k=8000, or about 30 words per cluster). Adding the clusters improves target and F-measure scores, although this result is not statistically significant. 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.

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, 2015) for overall F-measure; we note that our tasks differ as described in section 2.

Results on blind test Table 6 shows the results on unseen test data for best-linguistic using D3, D3+ATB and with clusters using k=8000. The results are similar to what was observed in the development data.

7.3 Error Analysis

We analyzed the output of our best linguistic models on the development set, and observed the following kind of errors:

Implicit Sentiment This was the most common kind of error observed. Commenters 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 post in question was using implicit language to express negative sentiment; the English gloss
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.

Table 7: Good and bad examples of output by SMARTies. Gold annotations for targets are provided in the text with ‘-’ or ‘+’ reflecting negative and positive sentiment towards targets.

| 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 7: Good and bad examples of output by SMARTies. Gold annotations for targets are provided in the text with ‘-’ or ‘+’ reflecting negative and positive sentiment towards targets.

is brains, which appears as a positive subjective word in the MPQA lexicon. The posts also contained cases of complex coreference where subjective statements were at long distances from the targets they discussed.

Annotation Errors Our models often correctly predicted targets with reasonable sentiment which were not marked as important targets by annotators; this points to the subjective nature of the task.

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.

Primary Targets The data contains multiple entity targets and not all are of equal importance. Out of the first 50 posts manually analyzed on the dev 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 correct targets (81.6%).

In general, our analysis showed that our system does well on posts where targets and subjective language are well formed, but that the important target identification task is difficult and made more complex by the long and repetitive nature of the posts. Table 7 shows two examples of the translated output of SMARTies, the first on more well-formed text and the second on text that is more difficult to parse.

8 Conclusions

We presented a linguistically inspired system that can recognize important entity targets along with sentiment in opinionated posts in Arabic. The targets can be any type of entity or event, and they are not known beforehand. 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. Our task is further complicated by the informal and very long sentences that are used in Arabic online posts. We showed that the choice of morphological representation significantly affects the performance of the target and sentiment models. This could shed light on further research in target-specific sentiment analysis for morphologically complex languages, an area little investigated previously. We also showed that the use of semantic clusters boosts performance for both target and sentiment identification. Furthermore, semantic clusters alone can achieve performance close to a more resource-rich linguistic model relying on syntax and sentiment lexicons, and would thus be a good approach for low-resource languages. Integrating different morphological preprocessing schemes along with clusters gives our best result.

Our code and data is publicly available. Future work will consider cross-lingual clusters and morphologically different languages.

Acknowledgments

This work was supported in part by grant NPRP 6-716-1-138 from the Qatar National Research Fund, by DARPA DEFT grant FA8750-12-2-0347 and by DARPA LORELEI grant HR0011-15-2-0041. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S government. We thank anonymous reviewers for their helpful comments. We thank Yves Petinot for providing feedback on the paper. We thank Nizar Habash and Mona Diab for helpful discussions.

6www.cs.columbia.edu/~noura/Resources.html
References


