Columbia_NLP: Sentiment Slot Filling

Sara Rosenthal
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
sara@cs.columbia.edu

Suvarna Bothe
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
suvarnabothe@gmail.com

Kathleen R. McKeown
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
kathy@cs.columbia.edu

Abstract

In this report, we present a follow up to our 2013 system which, given a named entity and the polarity (positive or negative) of an opinion expressed either by or towards it, finds all entities that could reasonably be the targets or holders, respectively, of that sentiment. The system operates in three steps: extracting viable entity pairs, analyzing the subjectivity of the text relating them, and classifying the polarity of the sentiment expressed. In addition, this year, we developed a sentiment knowledge base to be used as a prior sentiment scorer for each query entity.

1 Introduction

When text is presented to its audience as objective, expressions of sentiment are often elusive, obscured by evasive language or displaced from one entity to another. Criticism, for example, is often implicit, and its source difficult to locate:

> As the government wraps up its Troubled Asset Relief Program, the company that received the most from the fund, the American International Group, is offering an exit plan with no clear sense of whether the taxpayers will end up with a gain or a loss (*NYT*, January 10, 2010).

Here, does the reporter express opinion or fact? If it is sentiment, the target is unclear: perhaps the unaccountable financial firm, its government caretaker, or both. Indeed, opinion, though often difficult to read, permeates text like newswire, found in a candidate for political office criticizing an opponent, a victorious sports team congratulated by a rival coach, or an actor put down in a snide review. But in many cases, it can be difficult for even a human reader to determine who, exactly, the opinion comes from, and toward what, exactly, it is targeted.

Through our participation in the Sentiment Slot filling Task of TAC KBP 2014, we have developed a system to address the challenges of identifying pairs of entities related by sentimental expressions, both in newswire text and online discussion forums. In this paper, we briefly describe our previous system (Rosenthal et al., 2013) which identifies named entities and entity mentions, establishes grammatical relationships, and determines the polarity of any sentiment that may exist between them. This year we have additionally built a sentiment knowledge base which is used as a prior polarity score for the candidate entities for the query.

In the rest of this paper, we describe past work in sentiment detection, our procedure for tackling the unique challenges of slot filling with sentiment, including changes to the 2014 system, provide a brief error analysis, and our continued work in the area.

2 Related Work

There has been a large amount of work on sentiment detection. Wilson et al. (2005); Wiebe et al. (2005); Turney. (2002); Pang and Lee (2004); Beineke et al. (2004); Kim and Hovy. (2004); Agarwal et al. (2009), perform sentiment detection on edited text, while on the other hand, more recent work, Chesley et al. (2006); Godbole et al. (2007); Yu and Kübler (2011); Mei et al. (2007a); Go et al. (2009); Barbosa and Feng (2010); Bermingham and Smeaton (2010); Agarwal et al. (2011); Pak and Paroubek (2010); Rosenthal and McKeown (2013), gear their system towards social media, such as Weblogs and Twitter.

There has been previous work that focuses on sentiment towards an entity or topic. One such system is Godbole et al (2007) where they determine whether the sentiment towards an entity within a corpus is positive or negative and how it changes over time. Mei et al (2007), model sentiment towards the main topics in a document. Jiang et al (2011) perform sentiment towards a topic in Twitter. Nasukawa and Yi (2003) capture sentiment towards the topics in a document by exploring all entities where there is a semantic relationship. Similarly to our approach they explore the dependency between two entities to determine their semantic relationship.

Our supervised sentiment detection system builds off of an existing algorithm (Agarwal et al., 2009) developed for use on newswire documents and adapted to detect sentiment in social media (Rosenthal and McKeown, 2013; Rosenthal et al., 2014). The system predicts the polarity or subjectivity of a phrase in a given sentence (or tweet). The system initially uses lexical scoring to determine the polarity or subjectivity of a phrase using the Dictionary of Affect in Language (DAL) (Whissel, 1989) augmented with WordNet (Fellbaum, 1998). It then generates many features including lexical, syntactic, and stylistic features to automatically detect the subjectivity or polarity of the phrase. The system is described in further detail in 3.

Our system differs from previous systems in that it performs subjectivity and polarity detection at the phrase level using a system developed specifically for this purpose. This results in a more fine-grained prediction which is particularly beneficial in the TAC KBP 2014 evaluation, where the goal is to determine the sentiment between two entities as opposed to the entire document.

3 Method

Given a query, {**positive/negative** sentiment **from/towards** Entity **X**}, our system first finds all candidate entities occurring near the query entity. It then weeds out any entities that do not have a valid relationship. Afterwards, it determines whether the sentiment matches that in the query. All duplicate mentions that co-refer to the same entity (such as “Klein” and “he”, which refer to U.S. Representative Ron Klein) are winnowed into a single slot filler based on the confidence reported by the sentiment polarity analysis, and the most representative name is reported. We also experiment with using only the mention of an Entity that appears closest to the query. This can be useful as mentions occurring later tend to be less likely to be relevant based on distance from the query. Query examples are shown in Table 1. In the following sections, we describe each of the steps in greater detail.

3.1 TAC KBP 2013

Our main system techniques are identical to those developed for TAC 2013 (Rosenthal et al., 2013) and are discusses in more detail in Rosenthal et al (2013b). We used the SERIF co-reference/NER annotations provided by TAC KBP 2014 to obtain relevant entities for the evaluation. We only looked at entity mentions that were close in proximity to the query entity and its mentions. We then used the Stanford CoreNLP’s dependency parser to identify whether each entity acted as an object or a subject in relation to its surrounding text, and removed entity pairs that did not have the necessary grammatical relationships to each other to be expressing sentiment in the correct direction.

We perform sentiment detection using the system described in prior work (Rosenthal and McKeown, 2013; Rosenthal et al., 2014). The system pre-processes the sentences to add Part-of-Speech tags (POS) and chunk the sentences using the CRF tagger and chunker (Phan, 2006b; Phan, 2006a). It then applies the Dictionary of Affect and Language (DAL) (Whissel, 1989) augmented with WordNet (Fellbaum, 1998) to the pre-processed sentences. The DAL is an English language dictionary used to measure the emotional content of texts with scores for the pleasantness, activeness, and imagery. If a word was not found in the DAL, it was looked up in WordNet to locate synonyms and, barring their availability, hypernyms of words in the DAL. The scores were used to generate lexical-stylistic features (e.g. slang, hashtags, word lengthening, exclamation points), and lexical and syntactical features

Query (entity, sentiment)	**Text** (correct phrase labeled, entities bolded)	Valid Slotfillers	Invalid Slotfillers
Pelosi, pos-from	Indeed [**liberals** credit *Pelosi*] with pressuring **Obama** when **he** was inclined to cave.	liberals	"Pelosi" is not an object in relation to "Obama" and "he"; they are not evaluated for sentiment
Barber, pos-towards	[**Talib's** pick in the fourth quarter was about as clutch of a play that I've seen around here in a long, long time, *Ronde Barber*] said of his **fellow cornerback**.	Talib	Though "fellow cornerback" refers to Talib, this mention would be eliminated because "said of" is not subjective
Israel, neg-towards	"This assault proved once again, clearly, that the current [*government of Israel* does not want peace in the region," **Erdogan**] told reporters in **Chile**.	government of Israel	Text between "Erdogan" and "Chile" is not subjective
Benedict, neg-from	But U.S. [**victims** of clerical abuse were not impressed by **Benedict's**] selections, saying some of the **bishops** themselves had "troubling" records on confronting abuse.	victims	"Benedict" is not an object in relation to "bishops"

Table 1: Examples of queries, expressions of sentiment, and valid and invalid slot fillers

(e.g POS and n-grams for the target phrase and those surrounding it). These feature sets are reduced using chi-square in Weka (Hall et al., 2009). In addition, new to 2014 (Rosenthal et al., 2014), we included a feature related to the average SentiWordNet (Baccianella and Sebastiani,) score of the phrase which provided a small improvement (˜2%) to prior sentiment results. We run two variants of the phrase-based opinion detection system for this task. The first determines if a phrase is subjective, while the second classifies polarity.

3.2 Sentiment Knowledge Base

Any given entity will have a tendency towards one polarity based on the overall opinion towards that entity. For example, the overall opinion towards a murderer may be negative whereas the opinion towards a religious leader may be positive. We surmise that this general opinion can be used as prior knowledge to provide more weight towards the polarity expressed to the entity in a specific situation.

We build the knowledge base by finding mentions of each query entity throughout the corpus. We then find the polarity of the sentence the entity occurs in and add it to the knowledge base, storing the total sum of positive, negative, neutral, objective, and subjective occurrences. Examples from the knowledge base are shown in Table 2. For example, 63 occurrences of sentiment towards Bronislaw Komorowski were positive and 37 were negative indicating an overall positive opinion towards the entity. To improve speed we only took a maximum of 100 sentences per entity across the corpus. In addition we limited the amount of mentions per document to 5 exact matches of the entity. In the future, we would like to expand the knowledge base to include all entities in the corpus.

3.3 Confidence

In the overall KBP system, we filter out answers based on several metrics by reducing or increasing the overall confidence of the answer:

- Entity is a noun: $+.1$
- Entities have Subject/Object Relationship: $+.2$
- Entities have Object/Subject Relationship: $-.2$
- Prior Sentiment (in Knowledge Base) is the same as the Query Sentiment: $+.1$
- Author is the Entity and the word I is in the justification: $+.1$
- Author is the Entity and the word I is not in the justification: $-.1$

Entity	Positive	Negative	Neutral	Objective	Subjective
U.N. Watch	0	3	0	0	3
Bronislaw Komorowski	63	37	0	0	100
Fred Phelps	43	57	0	0	100
Morocco	62	40	0	0	102
Falungong	24	76	0	0	100
Luiz Inacio Lula da Silva	39	61	0	0	100

Table 2: Examples of Entities in the Knowledge Base

Error	Occurrence
Invalid Entity	14
Incorrect / Lack of Relationship	25
Incorrect Sentiment	6
Sentiment From a Place to the Query	8

Table 4: The common types of errors found by analyzing a subset of 8 queries consisting of 44 answers

4 Experiments and Results

Our system submission included 5 runs with varying combinations of filters and a confidence threshold, outlined in Table 3. We experiment with using or excluding the knowledge base (KB). We also experiment with using the mention closest (First Mention) to the query entity as the only candidate answer. Our best performing system was that which employed the most permeable filter combination (Run 5), allowing all entity mentions and no threshold, with an F-Score of 10.3%. Our runs show that the knowledge base caused a decrease in performance as evident in comparing runs 3 and 4. The change in confidence due to the knowledge base decreased the recall more than it increased the precision.

5 Discussion

It is clear from our results that the system did not perform well. We experimented with changing the threshold of the answers from run 5 and found that different variations did not cause the F-score to improve, but rather decrease as shown in the bottom half of Table 3. This indicates that changes to our confidence metrics are necessary for our system.

The common cause of error is due to choosing entities poorly. For example, given the query "positive sentiment towards Federer", our system chose the "French" as a candidate entity with the justification "Federer said the rivals chatted briefly Wednesday, and Nadal congratulated him for winning the French". There are several problems with this answer. First of all, places are in general not good entities. This is especially the case when trying to find sentiment *towards* a person as a place can not have an opinion about a person. The second issue with this answer is that SERIF did a poor job of extracting the entity as the entity should be the "French open". On a more positive note, the sentiment of the justification phrase was correctly determined to positive. Another issue is choosing entities that are not related to the query entity. For example, given the query "positive sentiment towards Mayor Bloomberg", our system chose "Bush" as the candidate entity with the justification "The mayor said that Obama deserved praise for working out a deal with Republican leaders to retain Bush". The error here is that there is no subject/object relationship between Bush and Bloomberg.

We analyzed two queries of each type (8 in total) for a combined total off 44 answers. The types of errors found are shown in Table 4. It is clear that the majority of errors are due to invalid entities and entity relationships indicating that improvement on top of SERIF and dependency parsing are necessary. The majority of the sentiment errors shown were related to one query where the justification was objective text describing the history of a person. Finally, we also computed the number of errors due to choosing a place as the candidate entity. This error can be avoided by excluding all places as answers to pos- and neg-towards person queries as a place will usually not have an opinion towards a person.

6 Conclusion and Future Work

Sentiment slot filling continues to remain a difficult task. Our error analysis indicates that more time

Run	First Mention	Threshold	Knowledge Base	Precision	Recall	F-Score
1	Yes	None	Yes	7.1%	17%	9.9%
2	Yes	None	No	7.1%	17%	9.9%
3	Yes	.75	Yes	9.8%	8.3%	9.0%
4	Yes	.75	No	9.6%	9.2%	9.5%
5	No	None	Yes	6.8%	20.7%	10.3%
5	No	.50	Yes	18.3%	6.6%	9.8%
5	No	.75	Yes	12.9%	8.5 %	10.2%

Table 3: The top half indicates runs submitted to TAC 2014. The bottom half refers to varying the threshold value for the results of run 5.

needs to be focused on finding correct entities and entity relationships rather than sentiment detection. In fact, it is likely that an excellent entity relationship system would do well at this task even if sentiment was ignored altogether. In the future, we would like to explore using additional coreference tools and taking a more thorough look at the grammatical parse trees to determine the proper relationship between two entities.

7 Acknowledgements

This paper is based upon work supported by the DARPA DEFT Program. 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.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen R. McKeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 24–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, June. Association for Computational Linguistics.

Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING (Posters)*, pages 36–44.

P. Beineke, T. Hastie, and S. Vaithyanathan. 2004. The sentimental factor: Improving review classification via human provided information. In *Proceedings of ACL*.

Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In Jimmy Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An, editors, *CIKM*, pages 1833–1836. ACM.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 27–29.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *ACL*, pages 151–160.

S. M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *In Coling*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007a. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of WWW*, pages 171–180, New York, NY, USA. ACM Press.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007b. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the Sixteenth International World Wide Web Conference*, WWW. World Wide Web Conference Committee.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 70–77, New York, NY, USA. ACM.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL.*

Xuan-Hieu Phan. 2006a. Crfchunker: Crf english phrase chunker.

Xuan-Hieu Phan. 2006b. Crftagger: Crf english phrase tagger.

Sara Rosenthal and Kathleen R. McKeown. 2013. Columbia nlp: Sentiment detection of subjective phrases in social media. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, Semeval, pages 478–482, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Rosenthal, Gregory J. Barber, and Kathleen McKeown. 2013. Columbia nlp: Sentiment slot filling. In *proceedings of the TAC KBP 2013 workshop*, Gaithersburg, Maryland, USA.

Sara Rosenthal, Kathy McKeown, and Apoorv Agarwal. 2014. Columbia nlp: Sentiment detection of sentences and subjective phrases in social media. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 198–202, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL.*

C. M. Whissel. 1989. The dictionary of affect in language. In *R. Plutchik and H. Kellerman, editors, Emotion: theory research and experience*, volume 4, London. Acad. Press.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.*

T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. In *Proceedings of ACL.*

Ning Yu and Sandra Kübler. 2011. Filling the gap: semi-supervised learning for opinion detection across domains. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 200–209, Stroudsburg, PA, USA. Association for Computational Linguistics.