

Columbia_NLP: Sentiment Slot Filling

Sara Rosenthal

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

Gregory J. Barber

Dept. of Computer Science
Columbia University
New York, NY 10027, USA
gjb2120@columbia.edu

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 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.

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 2013, 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 describe processes to identify named entities and entity mentions throughout a large corpus, establish the appropriate grammatical relationships between them, and determine the polarity of any sentiment that may exist between them.

It was unclear from the evaluation guidelines that the answer could be found anywhere in the corpus. Therefore our initial system analyzed only one document for each query, yielding poor results. But this has since been corrected, and the system now makes use of the full 2 million document corpus.

Because there is no standard list of slot fillers against which to compare our updated results, it is difficult to track the system's progress. However, we expect significant improvement in future slot filling evaluations. A manual evaluation of the slot fillers returned for 4 queries yielded an accuracy of 17.6%, versus 1.6% in the initial evaluation. Recall cannot be calculated accurately, as the total number of correct slot fillers is dependent on which slot fillers are found by the participating teams and, thus, evaluated for correctness by the evaluation organizers. However, because Columbia supplied a very small number of correct slot fillers to the initial evaluation key, we can expect that Columbia's results utilizing the

full corpus will constitute a larger part of an updated key, resulting in significantly improved recall.

In the rest of this paper, we describe past work in sentiment detection at both the phrase and document level, our procedure for tackling the unique challenges of slot filling with sentiment, including changes to the system after the TAC 2013 evaluation, 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); Birmingham 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). 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 2013 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. Query examples are shown in Table 1. In the following sections, we describe each of the steps in greater detail.

3.1 Establishing Entity Relationships

We used the SERIF co-reference/NER annotations provided by TAC KBP 2013 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 (i.e. for newswire, found within the same paragraph, or for discussion forum, found within the same post) throughout the document.

Next, we used 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. If the query entity expressed sentiment towards a target, for example, all pairs for which the query entity is an object in the sentence and a target entity is the sub-

Query (entity, sentiment)	Text (correct phrase labeled, entities bolded)	Valid Slotfillers	Invalid Slotfillers
Pelosi, pos-from	Indeed [liberals credit <i>Pelosi</i>] 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, <i>Ronde Barber</i>] 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 [<i>government of Israel</i> 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

ject were discarded; if the query entity is receiving sentiment from another entity, the pairs for which the query entity is a subject and the target an object are discarded.

For example, consider the following query.

Negative sentiment from Allan West

In the following text containing the query entity, potential target entities are bolded:

Klein said **he** doesn’t regret his votes for the health care and stimulus bills, measures that *West* and other **Republicans** used against **Democrats** nationwide.

Because “West” is a subject in the query, “Democrats” is filtered out as a possibility because it is acting as an object in the sentence. The other three entity mentions are kept for further analysis.

3.2 Sentiment Detection

We perform sentiment detection using the system described in prior work (Rosenthal and McKeown, 2013). 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 (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). We run two variants of the phrase-based opinion detection system for this task. The first determines if a phrase is subjective (a phrase can be labeled “subjective” or “objective), while the second classifies polarity (“positive” or “negative” if sentiment is present. For example:

[Klein said he doesn’t regret his votes for the health care and stimulus bills, measures that West]/**SUBJECTIVE** and other Republicans used against Democrats nationwide

If the phrase is evaluated as objective, the corresponding entity pair is not considered further. In the above sentence, for example, “Republicans” is removed as a possibility because the text between it and “West” (“and other”) is not a subjective phrase.

[Klein said he doesn't regret his votes for the health care and stimulus bills, measures that West]/**NEGATIVE** and other Republicans used against Democrats nationwide.

Thus, "Klein" and "he" are returned as potential slot fillers for the task. Additional examples are shown in Table 1.

4 Experiments and Results

We describe two sets of experiments: The submitted system (document based) and an improved system (corpus based). We present two systems because our initial system only explored the query document for candidate entities in contrast to the entire corpus. The only indication in the evaluation guidelines that the answers could be found any where in the corpus was in the example. Although it may be naturally assumed in the other SlotFilling evaluations that the answer could be found anywhere in the corpus, this is not generally the case in sentiment tasks. Being that this was our first time participating in TAC KBP, we assumed the evaluation was document based.

4.1 Document Based

Our initial system submission included 5 runs with varying combinations of filters and a confidence threshold, outlined in Table 2. These were performed within the query document only. This experiment yielded poor results because many queries had been included with the intention that systems would search thousands of documents in which the query entity appears. Our system returned NIL for many queries when searching only in the document. Thus, our best performing system was that which employed the most permeable filter combination (Run 5), assuming the subjectivity of all extracted phrases. The system returned a total of 9 correct slot fillers, resulting in an F-score of 1.2%.

4.2 Corpus Based

Following the initial submission, we retooled our system to search the full corpus for slot fillers, yielding improved results. The difference in scope is substantial; a figure such as Sasha Baron Cohen, for example, appears in 275 documents, while a geopolitical entity such as Uganda appears in over 3000.

Run	S/O Used	Subjectivity	Threshold
1	No	Yes	None
2	Yes	Yes	None
3	No	Yes	.75
4	Yes	Yes	.75
5	Yes	No	None

Table 2: Runs submitted to TAC 2013

In addition to the filters for objectivity and improper subject/object relationships, we also include a length filter that requires that entities appear within 150 characters of each other, increasing the likelihood that any sentiment identified in the intervening text relates the two entities.

The results of our four experiments, including the initial submission to TAC 2013, are summarized in Table 3.

For queries whose entities span thousands of documents, sentiment slot filling results are exceedingly difficult to check. Though the gold slot fillers provided by the TAC evaluators provide a good starting point to determine which results are correct, they do not cover all instances of sentiment in the large corpus of documents. Indeed, an analysis which excluded all slot fillers that did not come from documents appearing in the gold slot filler document covered only 225 of the slot fillers in our latest run, of which 42 were correct, and 183 incorrect. A further 1928 slot fillers returned by our system and which came from that document subset were not evaluated by TAC for correctness. Regardless, these preliminary results indicate a significant improvement from our precision measure on the initial submission, which increased from 1.6% to 18.7%.

For an accurate measure of precision, it is necessary to verify each slot filler by hand. This has been done for a total of 175 slot filling responses, from 4 separate queries. Precision on this narrow subset was 13.1%. Additionally, each slot filler was evaluated by three measures: coreference annotation, grammatical relationship between entities, and sentiment determination. The results of this evaluation are summarized in Table 4.

Coreference annotations, provided by SERIF, contribute significantly to incorrect slot filler determinations, particularly in discussion forums. For example, in a discussion forum about Adele's "curves"

Evaluation	Queries Tested	# Results	# Correct	# Incorrect	Precision	Recall	F-Score
Single Document	160	539	9	530	1.7	1.0	1.2
Limited Documents	160	225	42	183	18.7	4.1	6.7
All Documents	40	5566	12	5554	.2	3.1	.38
Hand-Annotated	4	175	23	152	13.1	N/A	N/A

Table 3: Summary of Results

in which the queried entity “Nicki Minaj” was briefly mentioned, the pronouns “she” and “her” were incorrectly attributed to Minaj when they should have referred to Adele, yielding 8 incorrect slot fillers on a single document. In another forum, the queried entity “David Frum” was incorrectly linked with the pronoun “I”, yielding 12 incorrect slot fillers. Overall, 29.2% of the returned slot fillers displayed some kind of coreference problem, including 43.2% of slot fillers returned from discussion forums (the discussion forum slot fillers were dominated by the two specific cases indicated above). Apart from incorrect coreference chains, problems included incomplete coreference names and coreference mentions that incorrectly referred back to the queried entity.

Issues with coreference annotations were the clearest source of error, and it is in general difficult to single out further sources of error, because incorrect sentiment detection and incorrect grammatical relation determinations almost always occur together. However, a number of conclusions could be drawn from the hand-annotated evaluation.

(1) **Polarity detection is our strongest filter:** The final filter of the system evaluated incorrectly in only 13.1% of the phrases examined.

(2) **Grammatical relation analysis can be improved:** Of the remaining incorrect slot fillers (i.e. those in which polarity determination or coreferences were not at fault), 27.7% were an explicit subject-object violation not picked up by the filter, while 15.4% had an extended, or indirect grammatical relationship with the query that was not an explicit violation. The remaining 56.9% displayed no relationship.

(3) **Subjectivity determination is difficult in newswire:** In the vast majority of cases, even if a grammatical relationship was incorrect, or there was no relationship found, the subjectivity filter should

Reason	Number	% of Results
Incorrect Coreference	52	29.7
No Relationship	37	21.1
Incorrect Relationship	18	10.3
Indirect Relationship	10	5.7
Subjectivity Incorrect	70	40.0
Polarity Incorrect	23	13.1

Table 4: Hand-Annotated Evaluation Summary: Reasons for Incorrect Slot Filler

have evaluated the text between the entities as “objective” and removed the slot filling entity from consideration. This difficulty is unsurprising, as newswire text, though generally written from an objective standpoint, often uses complex grammar and incorporates a highly subjective vocabulary, as in the example shown in the Introduction, or here, in a recent entertainment article from *The New York Times*:

If he never succeeds in diminishing her appeal, its both because Ms. Blanchett maintains a vise grip on the characters humanity and because it becomes inexorably clear that, while losing her money helped push Jasmine over the edge, it was also the dirty, easy money, what it promised and delivered, that drove her nuts to begin with.

For humans, as well as machines, classification can be difficult. The subjectivity filter encountered few issues with text extracted from the discussion forum, which is usually simple in structure and highly opinionated.

5 Conclusion and Future Work

The system described in this paper is an initial stab at detecting sentiment towards/from an entity. Sentiment slot filling continues to remain a difficult task,

layering three already challenging areas of study: subjectivity detection, sentiment polarity classification, and the recovery of entities with a set of valid grammatical relationships.

There are, however, a number of improvements to be made before further slot filling evaluations. We are looking into partnering with other researchers to develop an additional coreference system - either one which verifies or extends the SERIF annotations, or which replaces them completely. We will also experiment further with utilizing the grammatical parse trees of extracted phrases. There are instances (particularly in discussion forums where, unlike in newswire, the text is not conveniently divided into 1-2 sentence paragraphs), in which a more thorough look at the grammatical parse trees could help us determine when no relationship exists between two entities. As an intermediate step, it would be best to factor a lack of a grammatical relationship into our confidence determination.

6 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 Biadisy, 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.
- 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 Srinivasaiyah, 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. 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.
- 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.