SUNY Albany Sentiment System in TAC KBP Sentiment Slot Filling Evaluation 2014

Samira Shaikh[1], Tomek Strzalkowski[1,2], John Robert Giarrusso[1] and Veena Ravishankar[1]

[1]State University of New York
– University at Albany
samirashaikh@gmail.com

[2]Polish Academy of Sciences

tomek@albany.edu

Abstract

In this document, we detail our system that participated in TAC KBP 2014 Sentiment Slot Filling Evaluation. A two-pronged approach was developed to extract sentiment from two types of data genre provided in KBP Sentiment Slot Filling 2014 corpus – Newswire and Discussion Forums. We achieved an average F-score of 25% across five runs that were submitted for evaluation, which is twice the highest F-score achieved by teams that participated in the 2013 Sentiment Slot Filling Evaluation. In internal evaluations on training data, an F-score of 33.5% was achieved.

1 Introduction

The SUNY Albany Sentiment Extraction system participated in 2014 Sentiment Slot Filling Evaluation. Our approach towards sentiment is geared towards understanding sentiment of a speaker towards topics or entities in text.

2 Related Research

Affect in language is understood to mean the attitude toward a topic that a speaker/writer attempts to convey to the reader or audience via text or speech (van der Sluis and Mellish 2008).

There is a relatively large volume of research on sentiment analysis in language (Kim and Hovy, 2004; Strapparava and Mihalcea, 2007; Wiebe and Cardie, 2005; inter alia) that aim at detecting polarity of text. A number of systems were developed to automatically extract writer's sentiment towards specific products or services such as movies or hotels, from online reviews (e.g., Hu and Liu, 2004; Turney, 2002; Pang and Lee, 2008) or social media messages (e.g., Thelwall et al., 2010, Martineau and Finin, 2009). Socher et al. (2013) have recently used recursive neural tensor networks to classify sentences into positive/negative categories. Other relevant efforts in sentence level sentiment analysis include Sem-Eval Task[1].

In contrast, our objective is to isolate affect towards a given entity or topic – using the context it appears in as pieces of evidence that determine the affect polarity and strength.

3 Our Approach

In this section, we present the modules that participated in TAC KBP 2014 Sentiment Slot Filling Evaluation. We had two separate modules for the two types of data genre – Newswire and Discussion Forums.

For the Newswire type data, we used our Affect Calculus Module and for the Discussion Forum data, we used our Topical Positioning Module.

[1] https://www.cs.york.ac.uk/semeval-2013/task2/

3.1 Extracting Sentiment from Newswire data genre

The SUNY Albany system employs a novel approach to capture the sentiment of an entity towards another entity. We capture the contributing elements of sentiment, namely the sentiment holder, the sentiment target and the sentiment relation in an Affect Calculus. The Affect Calculus is then applied to determine the sentiment conveyed in text for the sentiment holder towards the sentiment target. The basic affect calculus is shown in Table 1 below.

We detail step-by-step the processing of a query for our algorithm

1. If the query is either pos-towards or neg-towards, we determine that the sentiment holder would be in an AGENTIVE role in the sentence. If the query is either pos-from or neg-from, then the sentiment holder would be in PATIENTIVE or PROPERTIVE role.
2. In the given document from the query id, we isolate all sentences that mention the query entity and parse them using Stanford Parser. We then isolate those sentences where the query entity is in the correct role (given step 1 above). For example, for a query entity to be in AGENTIVE role, it would have syntactic tag of SUBJ (and other variants). If the query is looking for sentiment from another entity, the name should be the object.
3. We then look for the sentiment *relation* that is the verb associated with the query entity and find a link to another entity that is syntactically linked to the verb. (This entity is denoted by X in the table above). If the relation to the subject is a variation of "says", "has", or "joined", then further processing is done in order to discover what it is the subject is "saying" or "joining" or "having" from the syntactic dependencies in the parse tree. If the relation is negated within the typed dependencies, this is accounted for as well.
4. If the sentiment relation has an affect score in the positive or negative spectrum of scores in our affect lexicon (depending upon the query – pos/neg), we use that the sentence and the associated sentiment target as potential outputs for sentiment slot filler. We use Affective Norms of Words (ANEW) lexicon (Bradley and Lang, 2010) to determine valence score. The algorithm's confidence scoring is also based on this valence score. The more intense the valence score, the higher the confidence.
5. The offset of the sentence is based on where the sentence currently being read is located within the article it was pulled from. In order to provide the evidence for answer to the query, 150 subsequent characters are pulled from the sentence. This creates a 150-character "window" which contains a section of the sentence where the relation and slot filler entity are found. The offset values for the sentence are updated to reflect this.

3.2 Extracting Sentiment from Discussion forum data genre

We used our Topical Positioning Module to provide answers to queries that had answers in Discussion Forum type of data. This module analyzes conversation to identify the salient topics of conversation – we call them meso-topics – and also identifies the polarity held by a speaker towards a given meso-topic. A certain number of subsequent mentions are required for a topic to be considered a meso-topic, which also leads to higher confidence about a speaker's attitude towards the topic. This constraint was relaxed during the evaluation, since the queries may sometimes be based on a single piece of evidence, not repeated mentions.

Relation type	**Type 1 (propertive)** *Rel(Target)*	**Type 2 (agentive)** *Rel (Target, X)*		**Type 3 (patientive)** *Rel(X, Target)*	
Relation/X		*X ≥ neutral*	*X < neutral*	*X ≥ neutral*	*X < neutral*
Positive	positive	positive	≤ unsymp	positive	≤ sympat
Negative	negative	≤ unsymp	≥ sympat	≤ sympat	≥ sympat
Neutral	neutral	neutral	≤ neutral	neutral	≤ neutral

Table 1. A simple affect calculus specifies affect polarity using a 5-point polarity scale [negative < unsympathetic < neutral < sympathetic < positive]. X is the second argument.

The module looks for polarized words within a 5-word window of a meso-topic. The valence score of words in the window are looked up in ANEW lexicon and an average score is taken. If the average score falls above the negative or positive threshold, we keep that utterance as potential answer to the slot filler.

In case of evaluation data, the meso-topic is the query entity. The speaker of the utterance is chosen as the sentiment holder. In case we determine the query is looking for reported sentiment, the utterance is sent to Affect Calculus module for further processing.

4 Description of Runs Submitted

For all of the run submissions, the run never accessed the web and the confidence values were based on the Affective Norms of Words (ANEW) lexicon valence scores. The higher the valence score in ANEW, the higher our confidence in the answer. The ANEW lexicon has valence score of words ranging from 1-9, 1 being more negative and 9 being more positive.

- Run 1. In the first run, there was no neutral range of scores and if the ANEW score was less than or equal to 5 and the query was looking for negative sentiment, then the answer was considered valid. If the score was greater than 5 and the query was looking for positive sentiment, then the answer was also considered valid.
- Run 2. In the second run, the same rules from run one apply, except that in the case where there were no "valid" answers because the algorithm rejected them, it would choose a slot filler entity as an answer anyway.
- Run 3. In the third run, the method was identical to run one, except we did not filter out non-named entities such as "this" and "that" as answers in the slot fillers.
- Run 4. In the fourth run the same rules apply as run one, except the neutral range is from 4.75 to 5.25 exclusive. So, if any relation in parse tree had a score that was in the neutral zone, it was rejected.
- Run 5. In the fifth run, the same rules apply as run one except the neutral range is from 4.00 to 6.00 exclusive. So if any relation in parse tree had a score that was in the neutral zone, it was rejected.

5 Evaluation and Results

On the training corpus, we calculated a recall value of 26.8%, a precision value of 44.8%, and an F-score value of 33.5%. The "tac_2013_kbp_sentiment_slot_filling_evaluation_annotations.tab" file was used to compare our answers with valid answers.

In Table 2, we show the performance of system achieved for the runs submitted in 2014 evaluation.

	Run 1	Run 2	Run 3	Run 4	Run 5
Precision	30%	27%	29.1%	28.6%	34%
Recall	22%	24%	23%	23%	18%
F-score	25.7%	25.3%	25.5%	25.5%	23.6%

Table 2. Performance across runs submitted by the suny_albany system during evaluation.

The best performing run was Run 3, with Run 4 a close second. In Run 5, we see that an increase in the range of neutral scores greatly added to our precision, but recall was negatively affected.

Experimenting with the range of neutral scores, so as to exclude spurious answers, while also including as many valid responses is a critical piece that we will work in the future.

We also determined that Inexact answers account for a sizeable number of responses from our system. An Inexact answer is one where the text spans (evidence) justify the found slot filler, but the slot filler includes only part of the correct answer or includes extraneous text. If we count the Inexact answers are Correct, we see an increase in system performance.

The updated results are shown in Table 3 below. An average F-score of 30% is achieved. These numbers are comparable to the performance seen on training data and compared against 2013 annotated data, where we achieved 33.5% F-score in our internal validations.

	Run 1	Run 2	Run 3	Run 4	Run 5
Precision	36.8%	33.4%	36%	34.5%	39.8%
Recall	27.2%	30%	27.7%	27.1%	21%
F-score	31.3%	32%	31.3%	30%	27.4%

Table 3. Performance across runs if Inexact answers are counted as Correct.

6 Discussion and Future Work

One clear piece of future work is to determine the best range of values to consider in the neutral zone from the range of valence scores in ANEW lexicon. Using an optimized range will maximize performance.

References

Margaret M. Bradley, and Peter Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-2. University of Florida, Gainesville, FL.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04.

Hu, M., and Liu, B. 2004. Mining opinion features in customer reviews. *In Proceedings of AAAI*, 755–760.

Martineau, J., & Finin, T. 2009. Delta tfidf: An improved feature space for sentiment analysis. *In Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*, 258-261.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. *Trends Inf. Retr*., 2(1-2):1–135, January.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng and Chris Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment TreebankIn *Proceedings Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013). Seattle, USA.

Carlo, Strapparava, and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Mike Thelwall, Kevan Buckley, and Georgios Patoglou. Sentiment in Twitter events. 2011. Journal of the American Society for Information Science and Technology, 62(2):406–418.

Peter D, Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 417–424.

Ielka van der Sluis, and C. Mellish 2008. Toward affective natural language deneration: Empirical investigations. affective language in human and machine. AISB 2008 Proceedings Volume 2.

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