English Opinion Analysis for NTCIR7 at POSTECH

Jungi Kim Hun-Young Jung Sang-Hyeob Nam Yeha Lee Jong-Hyeok Lee
Knowledge and Langauge Engineering Laboratory
Department of Computer Science and Engineering
Pohang University of Science and Technology
San 31, Hyoja-Dong, Nam-Gu, Pohang, Republic of Korea
{yangpa, blesshy, namsang, sion, jhlee}@postech.ac.kr

Abstract

We describe an opinion analysis system developed for Multilingual Opinion Analysis Task at NTCIR7. Given a topic and relevant newspaper articles, our system determines whether a sentence in the articles carries an opinion, if so, then extract the polarity and holder of the opinion. Our system uses subjectivity lexicons to score the sentiment weight of a word, in addition with a weight that reflects the discriminating power of the word. We borrow some techniques from Information Retrieval because discovering the importance and discriminating power of a word in a collection of documents is a commonly dealt issue in information retrieval tasks. We also use our own set of heuristics that are more specific to the task. Our system achieves high performance overall, with exceptional performances on polarity judgment of sentences.

Keywords: *Opinion Analysis, Multilingual Opinion Analysis Task, MOAT, NTCIR*

1 Introduction

Multilingual Opinion Analysis Task (MOAT) at NTCIR is a task of extracting opinions and related properties such as polarity, relevance to a topic, holder and target from a set of newspaper articles in English, Chinese, and Japanese. After the successful pilot workshop at NTCIR6 with participants from a number of research groups, MOAT at NTCIR7 called for finer granularity of analysis at sub-sentences (opinion-clauses) and an additional job of finding opinion targets [2].

Among the tasks defined for NTCIR7, We performed the judgment of opinion and polarity of sentences, and extracting holders of such sentences. Participating for the first time, our aim for NTCIR7 was to develop an initial system that performs reasonably well but more importantly has rooms for implementing different ideas for this task and in the future work. While most previous work focused on analyzing opinion-related properties, our work explores the usefulness of term weighting scheme in opinion analysis tasks.

Our system takes the form of a general lexicon-based opinion identification system, consisting of an opinion identifier, a polarity identifier, and an opinion holder extractor. The opinion analysis system utilizes various lexicons; opinion and polarity identifiers use a sentiment lexicon and a list of appraisal verbs to distinguish words containing sentiments, and our opinion holder extractor additionally requires a list of communication verbs for detecting entities expressing opinions in sentences.

Unlike previous work, we have attempted to explore the idea of weighting the informativeness of words; whether the appearance of the word is significant statistically, syntactically, or topically. Also, we consider the prior probability of a sentence being opinionated according to its context, namely the document it belongs. Term weighting and document smoothing are extensively studied subjects in information retrieval (IR). We have taken a few IR approaches to term weighting and sentence smoothing, as well as our own heuristics to reflect our first hypothesis described in section 2.

2 Hypotheses

Our work distinguishes itself from the previous work by grounding its basis on the hypotheses described below:

1. Opinionativeness of a word consists of a sentiment and an informative aspects, each represented by some measures independent of each other.

2. A sentence of a document with many opinionated sentences is more likely to be opinionated. Similarly, a document tends to contain either mostly

of positive sentences or mostly of negative sentences.

The first hypothesis emphasizes that words differ not only in opinion-related properties such as polarity and strength, but also differs in reliabilities of such properties in the language usage or in different contexts. Previous work has mostly dealt with learning the polarity and strength of words' or phrasal expressions' opinionatedness. However, such approaches fail to differentiate the discriminating power of terms, and adjust opinion strengths accordingly. The importance or dicriminating power of a word is not the strength of opinion. Rather, it is measure of how dicerning the word is compared to other words in the collection or the context of documents.

Secondly, we assume that a document consists either mostly of opinionated sentences or mostly of non-opinionated sentences, hence there is a prior probability of a sentence being opinionated hinted by the document it belongs. Here we consider the nature of the newspaper articles, if not any piece of writing, that a document on certain topics usually serves a purpose of either providing objective information or advocating one's opinions. Assuming that this proposition holds, a sentence will tend to follow the document's overall tendency of whether to express opinions or not.

3 Proposed System

Based on the hypotheses and the approaches commonly used in the previous work, we develop an opinion analysis system capable of detecting opinion, polarity, and opinion holder. We hypothized that the opinionatedness of each word consists of opinion-related properties represented as a sentiment weight and informativeness-related properties represented as a term weight (hypothesis 1).

Our system employs simple score functions to evaluate the opinionatedness or the polarity. As shown in the equations 1, an opinion score, and positive and negative polarity scores of a sentence are regarded simply as sums of products of the sentiment weight and the term weight of all words in the sentence.

$$\begin{aligned} Op(s) &= \sum_{w \in s} W_{Sentiment}(w) \cdot W_{term}(w) \\ Pos(s) &= \sum_{w \in s} W_{Sentiment_{pos}}(w) \cdot W_{term}(w) \\ Neg(s) &= \sum_{w \in s} W_{Sentiment_{neg}}(w) \cdot W_{term}(w) \end{aligned} \quad (1)$$

These equations are used as baselines to judge opinion and polarity of a sentence in our system. We describe in the following sections various terms used in the equations and introduce a a slightly modified version that reflects the hypothesis 2.

3.1 Sentiment Weght

To assess the sentiment weight of each word, we used SentiWordNet[1] [3] and a list of Appraisal Verbs[2] [12].

SentiWordNet is a set of WordNet synsets with automatically assigned positive, negative, and neutral probability scores. In our experiments, we have treated each word in WordNet synsets independently and assigned the scores of the synset it belongs. A word with different senses has multiple candidates of sentiment scores. In such cases, we choose the maximum scores.

Appraisal verbs are appraisal words from Levin's Verb Classes [17]. Unlike SentiWordNet, words in the appraisal verbs list are hand-picked, hence we consider them more reliable. We augment the sentiment weight of a word by adding a constant if the word exists in the appraisal verb list.

We simply summed positive and negative scores to compute the sentiment scores. According to the SentiWordNet, a subjectivity score of a synset is the sum of a positive score and a negative score of the synset. However, choosing the larger value of the two scores is also feasible because a word in a context is carrying a sentiment of either positive or negative opinion.

$$\begin{aligned} W_{Sentiment_{pos}}(w) &= SWN_{pos}(w) + Appraisal(w) \\ W_{Sentiment_{Neg}}(w) &= SWN_{neg}(w) + Appraisal(w) \\ W_{Sentiment}(w) &= SWN_{pos}(w) + SWN_{neg}(w) \\ &\quad + Appraisal(w) \end{aligned} \quad (2)$$

$$Appraisal(w) = \begin{cases} C & \text{if w is an appraisal verb} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Constant C is arbitrarily set to 1.5.

3.2 Term Weight

A word may have different informativeness in sentences according to its statistics in the document collection, the role in the sentence, or proximity to topical words.

We have defined the term weight of a word with three different factors that could affect the informativeness of a word.

$$\begin{aligned} W_{term}(w) = \; & W_{BM25}(w) \cdot W_{TreeDepth}(w) \\ & \cdot W_{TopicProximity}(w) \end{aligned} \quad (4)$$

[1] http://sentiwordnet.isti.cnr.it/

[2] http://lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz

Each term in the equation is described in the following subsections.

3.2.1 BM25 Retrieval Model

The issue of determining term weights according to its importance and discriminating power using statistical knowledge from a document collection has been extensively studied in the field of IR. In classical term-weighting model TF-IDF, words are weighted in the aspects of: (1) inter-document discriminative power (inverse document frequency, *IDF*) and (2) intra-document significance (term frequency, *TF*).

We naively take an information retrieval model popularly used in document retrieval systems. From variant versions of Okapi BM25, We implemented one as follows.

$$W_{BM25}(w) = \log \frac{N-df+0.5}{df+0.5} \cdot \frac{tf \cdot (k_1+1)}{tf+k_1 \cdot (1-b+b \cdot \frac{dl}{avgdl})} \quad (5)$$

Model parameters k_1 and b are set to $k_1 = 2.0$, $b = 0.75$, as is commonly done in the ad-hoc document retrieval. Term statistics are figured out from appropriate sources: term frequency (*tf*) and document length (*dl*) are computed using the test document, and document frequency (*df*) and average document length (*avgdl*) are estimnated from the NTCIR CLIR English newspaper corpus.

3.2.2 Depth in Dependency Tree

Dependency tree reveals the dominant and dependant relations among words in a sentence, such that a dominant word is the parent of dependant words forming a tree structure. In such tree graphs, the word of a parent node is described by the words of its descendant nodes. Therefore, as the height of a node in the dependency tree increases, the influence of the nodes on whole sentence decreases: the root with the most importance and the leaf nodes the least. We assume that opinionatedness of a node also decreases as the node is deeper in the tree.

A heuristic described in equation 6 assigns lesser weight to the nodes located deeper in the dependency tree. This simple rule assigns less weights to nouns or adverbs than to verbs, less weights to adjectives than to nouns, within a simple sentence, and less weights to subordinate sentences and clauses than main sentences and clauses.

$$W_{TreeDepth}(w) = DepTreeDepth(w)^{penalty} \quad (6)$$

The penalty factor of 0.9 is set arbitrarily without any tuning on a training corpus.

3.2.3 Topical Proximity

The object of the opinion analysis system is to find opinionated sentences from a set of relevant documents to a given topic. The heuristic in equation 7 boosts term weights of the opinionated words located near topical words, in the hope of rewarding the opinions about the topic.

If an opinionated word appears near topical words (*nouns*, *verbs*, *adjectives*, *adverbs*) in less than 2 traverse in a dependency tree, then its term weight is increased by 50%.

$$W_{TopicProximity}(w) = \begin{cases} 1.5 & \text{if distance to topic in} \\ & \text{dependency tree} <= 2 \\ 1.0 & \text{otherwise} \end{cases} \quad (7)$$

3.3 Opinion Prior

From hypothesis 2, we assume that sentences have prior opinion or polarity scores provided by the document it belongs. We use the Jelinek-Mercer method, a simple interpolation smoothing, to merge the score of a sentence with the score of the document. (equation 8)

$$\begin{aligned} Op_{Smooth}(s) &= \lambda \cdot Op(s) + (1-\lambda) \cdot \frac{\sum_{s' \in D} Op(s')}{|D|} \\ Pos_{Smooth}(s) &= \lambda \cdot Pos(s) + (1-\lambda) \cdot \frac{\sum_{s' \in D} Pos(s')}{|D|} \\ Neg_{Smooth}(s) &= \lambda \cdot Neg(s) + (1-\lambda) \cdot \frac{\sum_{s' \in D} Neg(s')}{|D|} \end{aligned} \quad (8)$$

3.4 Opinion Judgment

The opinion judgment of sentences are carried out using the equations 1 and 8 with different combinations of their sub-components , and a threshold value θ_{op}, which determines the minimal value of opinion scores $Op(s)$ and $Op_{Smooth}(s)$ to be judged as *opinionated*. We optimized θ_{op} using the NTCIR6 MOAT corpus, and the final value was tuned on the NTCIR7 MOAT Example corpus.

3.5 Polarity Judgment

Once the opinionatedness of a sentence is judged as opinionated, sentence polarity is determined using $Pos(s)$ and $Neg(s)$ in equations 1 and 8. If $Pos_{Smooth}(s)$ is greater than $Neg_{Smooth}(s) + \theta_{pol}$, then the sentence s is judged as *positive* (**POS**). If $Neg_{Smooth}(s)$ is greater than $Pos_{Smooth}(s) + \theta_{pol}$, then the sentence s is judged as *negative* (**NEG**). Otherwise, the sentence s is *neural* (**NEU**). The value of θ_{pol} is set to 0.1 arbitrarily.

3.6 Opinion Holder Extractor

To extract opinion holders, we exploited a set of communications and appraisal verbs, SentiWordNet, a named entities recognizer, and a syntactic parser. The list of communication and appraisal verbs are from [12]. We use the Stanford statistical parser [3] [15] to obtain dependency parses of English sentences, and the Stanford Named Entity Recognizer [4] [16] to recognize named entities in the sentences. We also manually compiled a list of non-named entity opinion holder candidates such as pronouns and professions found in the NTCIR6 English MOAT corpus.

We created a set of tuples containing words and its sentiment score (maximum of its positive and negative sentiment scores) in the SentiWordNet. Then, scores of all communication verbs were set to 0.9 and appraisal words 0.7.

Given a sentence, we find the most opinionated word using the compiled lexicon. From such word in the dependency tree, we traverse up the tree to its first ancestor node with POS as *verb*. We extract the *nominal subject* (n_subj) of the verb as the holder of the opinion expressed in the sentence. If a subject is not found, then "author" is set as the opinion holder of the sentence. If a subject is found, then from the NP chunk, we extract any named entities or opinion holder candidates are extracted as the opinion holder. If no named entity or opinion holder candidate is found, then we set the holder as the "author" of the document.

Regardless of the previous step, if a sentence includes quotation marks, then the speaker of the quote is extracted as the opinion holder of the sentence.

4 Experimental Results and Discussion

4.1 NTCIR6

We report in table 3.6 the best performance of our system tuned for NTCIR6, tuning the parameters λ for Jelinek-Mercer smoothing and θ_{op} for threshold. A system using only sentiment weights from SentiWordNet is set as our baseline, and we have set up different systems by adding different components to it. Emperically, we have shown that the idea for each components have worked, some with very exceptional improvements while most only mild. The system performs the best when every suggested components are used together, improving precision and f-measure exceptionally, but with slight loss in recall, over the baseline. Our proposed ideas worked particularly well for the polarity judgment tasks and has shown improvement of 35.5% and 25.4% in precision and f-measure, respectively, in lenient evalution scheme. The best performance of opinionated judgment of our system on the NTCIR6 English MOAT is comparable to the best systems at NTCIR6, and the perfrmance polarity judgment of our system out-performs NTCIR6's best system on polarity jugdment [1].

The performance of opinion holder extraction cannot be exactly measured because human intervention is required to judge partially matched answers. We report the performance in the worse case by answering "No" to every partially-matched answers, and in case using appropriate judgment of one of the authors.

4.2 NTCIR7

Using the best system validated on the NTCIR6 English MOAT corpus, we submitted three different runs KLE 1∼3, where λ and θ_{op} are optimized on F-measure, precision, and recall, respectively, using the NTCIR7 English MOAT Example corpus.

Our systems performed as expected from the results of test runs on the NTCIR6 corpus, ranking high in opinionated judgment tasks and out-performing other systems on polarity judgment tasks [1].

5 Conclusion

Despite the simple approach of our systems in using a simple tf-idf-like factor and some heuristics to reflect term weighting into the system of analysing opinions in newspaper articles, overall, it has achieved high performance in opinionated and polarity jugdment tasks. Also, our opinion holder extraction scheme has worked quite well, despite of its simple approach in finding the most opinionated words and their verbs, and extracting the verb's nominal subjects.

Although our system needs further improvements by firmly stamping its theoretical foundations, it has shown its potentials through emperical evaluations against various systems at NTCIR7.

Our future work includes strengthening the system's theoretical foundation based on the hypotheses that we have built our system upon.

Acknowledgments

This work was supported in part by MKE & IITA through IT Leading R&D Support Project and also in part by the BK 21 Project in 2008.

References

[1] Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi Chen, Noriko kando, and Chin-Yew Lin. Overview of Opinion Analysis Pilot Task at NTCIR-6. In *Proceedings of the sixth NTCIR Workshop*.

[3] http://nlp.stanford.edu/software/lex-parser.shtml
[4] http://nlp.stanford.edu/software/CRF-NER.shtml

Table 1. Performance of Opinion Analysis System on NTCIR6 Collection. Systems are optimized for F-measure using the NTCIR6 English corpus and lenient evaluations.

System	L/S	Opinionated			Polarity		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
SentiWN (2.2)	L	0.285	0.809	0.422	0.107	0.400	0.169
SentiWN+AppraisalVerb (2.9)	L	0.305	0.707	0.426	0.115	0.350	0.173
SentiWN+BM25 (7.2)	L	0.317	0.776	0.450	0.114	0.366	0.173
SentiWN+TreeDepth (2.0)	L	0.299	0.741	0.426	0.114	0.372	0.175
SentiWN+TopicProximity (2.5)	L	0.281	0.835	0.421	0.107	0.418	0.170
SentiWN+Smoothing (2.9/0.4)	L	0.296	0.783	0.430	0.111	0.387	0.173
All (7.2/0.4)	L	0.345	0.717	0.466	0.145	0.395	0.212
SentiWN (4.3)	S	0.071	0.444	0.122	0.030	0.264	0.054
SentiWN+AppraisalVerb (4.3)	S	0.071	0.495	0.124	0.030	0.294	0.055
SentiWN+BM25 (7.1)	S	0.064	0.791	0.118	0.027	0.467	0.051
SentiWN+TreeDepth (3.3)	S	0.071	0.404	0.120	0.032	0.259	0.057
SentiWN+TopicProximity (4.3)	S	0.069	0.448	0.119	0.029	0.269	0.053
SentiWN+Smoothing (4.5/0.5)	S	0.083	0.307	0.130	0.039	0.203	0.065
All (7.9/0.4)	S	0.073	0.592	0.131	0.038	0.437	0.071

Table 2. Performance of Opinion Holder Extraction System on NTCIR6 Collection.

L/S	Human Judgment	Precision	Recall	F-Measure
L	“No” to all questions	0.180	0.386	0.240
L	Authors’ judgment	0.202	0.433	0.276
S	“No” to all questions	0.035	0.379	0.065
S	authors’ judgment	0.039	0.418	0.072

Table 3. Performance of Opinion Analysis System on NTCIR7 Collection. KLE1 optimized for F-measure, KLE2 for precision, and KLE3 for recall, using the NTCIR7 MOAT English Example corpus and lenient evaluations.

System	L/S	Opinionated			Polarity		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
KLE1	L	0.353	0.727	0.475	0.155	0.422	0.226
KLE2	L	0.375	0.541	0.443	0.161	0.307	0.211
KLE3	L	0.274	0.933	0.423	0.122	0.552	0.200
KLE1	S	0.111	0.768	0.194	0.041	0.500	0.075
KLE2	S	0.119	0.579	0.198	0.042	0.357	0.074
KLE3	S	0.081	0.926	0.149	0.033	0.670	0.063

Table 4. Performance of Opinion Holder Extraction System on NTCIR7 Collection.

System	L/S	Precision	Recall	F-Measure
KLE1	L	0.400	0.508	0.447
KLE1	S	0.133	0.532	0.213

[2] Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In *Proceedings of the seventh NTCIR Workshop*.

[3] Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LERC'06)*, pages 417–422, Geneva, IT.

[4] Yi Hu, Jianyong Duan, Xiaoming Chen, Bingzhen Pei, and Ruzhan Lu. 2005. A new method for sentiment classification in text retrieval. In *Proceedings of the IJCNLP 2005*.

[5] Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 1367–1373, Geneva, CH.

[6] Soo-Min Kim and Eduard Hovy. 2006. Identifying and analyzing judgment opinions. In *Proceedings of HLT/NAACL*, 2006.

[7] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning Multilingual Subjective Language via Cross-Lingual Projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*.

[8] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer A Computer Approach to Content Analysis.*, MIT Press, Cambridge, MA.

[9] Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting emotional polarity of words using spin model. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140, Ann Arbor, US.

[10] Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2006. Latent variable models for semantic orientations of phrases. In *Proceedings of the EACL 2006*.

[11] Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

[12] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowlwdge Management (CIKM'05)*, pages 625–631, Bremen, DE.

[13] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT-EMNLP-2005*.

[14] Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the EMNLP 2003*.

[15] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

[16] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.

[17] Beth levin. English Verb Classes and Alternations: a preliminary investigation. University of Chicago Press, Chicago and London, 1993.