Analysis of the Evaluation Results for our Tasks in COAE2009

CHEN Mosha[1], WANG Rui[2], ZHANG Xiaojun[1], QIU Wei[1], ZHANG Yi[3], LI Tingyu[1], ZHANG Wenbo[1], and YAO Tianfang[1]

[1]UdS-SJTU Joint Research Lab for Language Technology
Dept. of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai 200240, China.
[2]Computational Linguistics Department, Saarland University, Germany
[3]DFKI GmbH, Saarbruecken, Germany
[1]{mosha,jiuren,hh190,hellowenniu}@sjtu.edu.cn;
[1]Xiaojun.zhang.iiken@gmail.com;
[1]yao-tf@cs.sjtu.edu.cn;
[2]rwang@coli.uni-sb.de;
[3]yzhang@coli.uni-sb.de;

Abstract

COAE2009 has five tasks and we take part in Task 3, 4 and 5. Task3 is designed for identification of the opinioned sentence; Task4 is designed for topic identification based on the sentences from Task3 and makes the polarity classification; Task5 is about opinion retrieval plus the sentiment polarity analysis. This paper will present our methods in the three tasks and finally draw our conclusion and present our future work.

1 Introduction

Text Opinion Analysis is a task of growing interest in recent years. Many researches on this issue begin to exist in top conference such as ACL, SIGIR and WI. Also international evaluation contest like Trec Blog Track and NTCIR begin addressing this issue in recently years. It is a relative new topic in Chinese language processing. Following last year's evaluation contest (COAE2008), the Chinese Information Processing Society of China holds the 2nd evaluation contest (COAE2009). COAE2009 has 5 tasks, focusing on sentiment classification, opinion sentence selection, topic extraction and topic retrieval. These tasks range from basic word level to complex chapter level. In COAE2008, we take part in the first four tasks and get good results, and this time we have Task 3, 4 and 5. Task3 is to recognize the opinionate sentences; Task4 is to identify the topic and then classify its polarity; Task5 is mainly about opinion retrieval and on this basis analyze the sentimental polarity.

The rest of this paper is organized as follows: Section 2 describes Task3; Section 3 describes Task4; Section 5 describes Task5; Section 6 gives the conclusion and future work.

2 Task3

In task 3, it is required to automatically identify 1000 opinionated sentences in test set Dataset1. That is, extract 1000 sentences that contain explicit sentiment polarity towards some point of view. The output of the result should be sorted by confidence. Meanwhile, the format is constraint by adding the participants' information and the number of article in Dataset1.

2.1 Problem Analysis

In Liu Bing's overview of opinion mining tutorial (Bing Liu, 2005). The main tasks in opinionated text analysis are consisted with the following ones: (1) Detect the sentiment element in documents. (2) Identify the polarity and the strength of sentiment element. (3) illustrate the relation between opinion object and sentiment element.

Engstrom (Engstrom , 2004) studied how the topic dependence affacts the accuracy of sentiment classification value is observed for a given statement. Nasukawa and Yi (Nasukawa and Yi, 2003) extracted positive or negative expressions on a given product name using handmade lexicons.

These opinion-mining issues are all based on the extraction of opinionated sentences in scaled texts. Therefore, how to identify the sentence with opinion works as a key role in sentiment analysis. In this task, we design two different algorithms to extract three types of sentences, as follows:

1) Sentence with explicit sentiment element, which are mainly sentiment adjective and adverb.

"炫目的色彩，动听的音乐，逼真的音效……这些都是张艺谋的长处。"

Comparatives (Doran et al., 1994)

Met-linguistic Comparatives: Those which compare the extent to which an entity has one property to a greater or less extent than another property

"与其说生气,罗纳尔多更多的是沮丧."

Propositional Comparatives: Those that make a comparison between two propositions. This category has subcategories:

Nominal Comparatives: They compare the cardinality of two sets of entities denoted by nominal phrases.

"保尔吃的香蕉比苹果多."

Adjectival Comparatives: In general, these comparatives appears with some comparative adverbs such as "更\更加\最"

"首先，它是目前备有 3 倍光学变焦 200 万像素数码相机中最薄最扁以及最轻的一款."

Adverbial Comparatives: They are similar to nominal and adjectival ones, instead of comparative adjective, they use adverbs as the description for certain properties:

"宝马 Z4 跑车比其他系列启动更迅速"

Sentences with explicit words or phrases which following sentiment/opinionated clauses. These words/phrases can be "认为", "觉得","指出", etc.

2.2 Solution

2.2.1 Linear combination based on sentiment element extraction

In this approach, we extract the adjectives and adverbs for each sentence. These words highly represent the sentiment polarity, and each of them obtains a polarity value, which describes the sentiment strength. The sentiment strength for a sentence is linear combined by all the opinionated elements. Considering the longer sentence may contain more sentiment elements, we normalize it by dividing. The following formula describes the strength value of sentence S:

$$Value(S) = \frac{\Sigma Value(w_i)}{|\text{Length of S}|}, \text{for all } w_i \in S$$

2.2.2 Classification base on word and phrase level

Firstly, we collect a corpus that is out of the test Dataset1. We annotate them by tagging the sentiment polarity, and build a two class's classifier after training process. By using this classifier, we could classify the test data set, and collect the result by sorting the confidence value. In our approach, we use Support Vector Machine for training and test. In the training step, three levels of feature sets are extracted automatically and manually, as follows:

a. Word level: in this level, a sentiment dictionary is used to match the adjective and adverb appears in target sentence.

b. Phrase level: we use Stanford Log-linear Part-Of-Speech Tagger (Toutanova, 2003) to annotate the corpus, and manually filter the POS templates which indicate the sentiment polarity.

c. Further more, we combine the word and the phrase template as a more specific feature.

2.3 Experiment and Results

In the above two solutions, we use the sentiment dictionary (Fang and Yao, 2008). Only 500 adjective and adverbs are selected by their higher confidence.

In solution 2, instead of features of sentiment dictionary, we manually add comparative adverbs and other verbs that introduce an opinionated clause. In total, the quantity of feature set is 512.

The Table 1 describes the comparison of our result and other competitors. In the first run, we simply use solution 1.The result is similar with the run 2, which used the feature set by words and phrase templates. But there is significant improvement when combine them together. This indicates the constraint of both words and its corresponding phrase template will help to identify the sentiment sentence in text.

Table 1 Result of Task 3

Run-tag	P@1000	Precision	Recall	F1	R-accuracy
Run1	0.402	0.40321	0.0603604	0.105002	0.0603604
Run2	0.418	0.419258	0.0627628	0.109181	0.0627628
Run3	0.461	0.462387	0.0692192	0.120413	0.0692192
MEDIAN	0.45	0.45	0.0675676	0.117493	0.0675676
MAX	0.625	0.625	0.0938438	0.163185	0.0938438

3 Task4

For the evaluation task 4, we identify the opinion objects from the subjective sentences, and classify the opinion polarities. In brief, we firstly take emotional words as cue for selecting subjective sentences, then we apply a log-linear model to rank all the candidate targets (i.e. the object of the opinion) together with their polarities, and finally we pick up the best target-polarity pair as the output.

Before extracting the features for the log-linear ranking model, we preprocess the corpus using a pipeline system, including the following modules in order,

Sentence boundary detection (with our own script based on regular expressions)

Word segmentation (Stanford Chinese Word Segmenter – Tseng et al., 2005)[1]

Part-of-Speech tagging (Stanford Log-linear Part-Of-Speech Tagger – Toutanova et al., 2003)[2]

Dependency parsing (MSTParser – McDonald et al., 2005)[3]

Semantic role labeling (our own system – Zhang et al., 2009)

The last module, semantic role labeler, is a type of shallow semantic processing technique, which normally reveals the predicate-argument relations between words or constituents in the sentence. For this task, we use the Chinese semantic role labeler described in (Zhang et al., 2009) to process all the documents provided by the evaluation task. The SRL system was trained on the Chinese PropBank and successfully participated in the CoNLL Shared Task 2009 (Hajic et al., 2009). Annotations in Chinese PropBank use role names like “A0, A1” to denote arguments, and “TMP, LOC, ADV” to identify temporal, location, and adverbial modification relations.

The main system starts from an existing emotional word dictionary (Liu et al., 2008), and use those words with strong polarities (±3, the strongest) as cue for selecting sentences from the whole corpus. In practice, we choose 7485 sentences as our subjective sentences, as well as the candidates for opinion object identification.

In order to build a supervised learning model, we manually go through about 1000 sentences and annotate 294 positive instances and 244 negative ones. The annotation labels we use are +*/-* for emotional words and +#/-# for opinion objects. Due to the limited size of the training set, we prune

1 http://nlp.stanford.edu/software/segmenter.shtml

2 http://nlp.stanford.edu/software/tagger.shtml

3 http://sourceforge.net/projects/mstparser/

the search space by only taking consideration on noun phrases as targets and the annotation is also restricted on the head word instead of the whole noun phrase.

With the manually annotated opinion objects and polarities, we developed two statistical classifiers to i) identify the opinion objects in the given subjective sentence, ii) classify the polarity of the opinion. Both classifiers are trained on the manually annotated dataset.

For the object identifier, the system starts from the emotion word, and search in the syntactic dependency graph for a potential path that can lead to a candidate opinion object. The system encode the dependency path, along with other surface, syntactic and semantic features (including words, part-of-speech, position in the sentence, dependencies related to either the object or the emotion word, semantic role relation between the emotion word and the candidate object, etc.). We adopt a log-linear model as follows,

$$P(t_i \mid e,S) = \frac{\exp(\sum_j \lambda_j f_j(t_i,e,S))}{\sum_{i'} \exp(\sum_j \lambda_j f_j(t_{i'},e,S))}$$

where e is the emotion word in sentence S, and t_i is the i^{th} candidate object. f_j is the j^{th} feature, and λ^j is its corresponding weight. The conditional probability $P(t_i \mid e,S)$ is calculated for each candidate object t_i in the sentences, and is passed to the next stage of processing.

In the opinion polarity classification stage, a similar log-linear model was defined for the conditional probability $P(t_i, p_j \mid e,S)$ where p_j is the polarity (positive or negative) of the opinion given a sentence S, emotion word e and object t. The feature set used in this step is similar to the ones used in the previous object identification step.

By multiplying the above two probabilities, we have the joint conditional probability of an

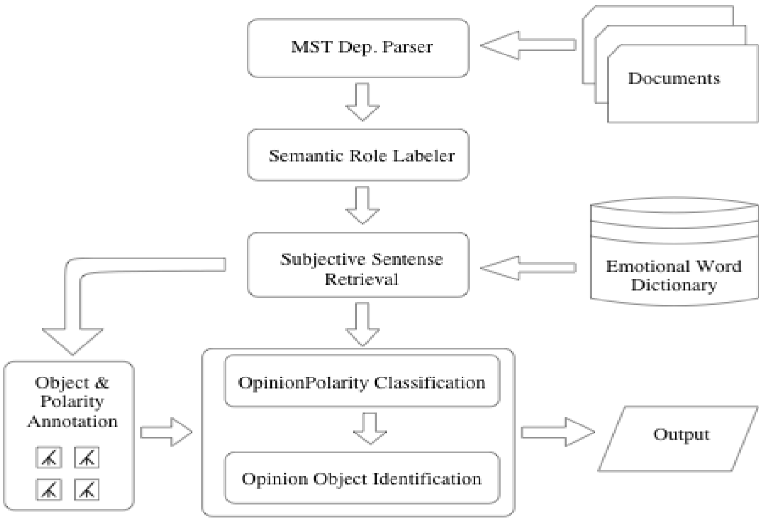

Figure 1 The Architecture of the System

object having particular opinion polarity given the emotion word in a sentence,

$$P(t_i, p_j \mid e,S) = P(t_i \mid e,S) \cdot P(p_j \mid t_i,e,S)$$

and we can simply pick the best pair of object and polarity (t,p),

$$(t,p) = \underset{(t_i, p_j)}{\operatorname{argmax}}(P(t_i, p_j \mid e,S))$$

After obtaining the (best) target word, we use the two rules to restore the whole noun phrase (i.e. the opinion object): 1) use the whole sub tree of the predicted word on the dependency tree; and if the resulting phrase is too short (less than two words) or too long (more than eight words), we 2) use the adjacent words with POS tags NR, PN, and NN[4] to concatenate them with the target word.

Finally, we take the top 1000 predicted target-polarity pairs (ordered by their probability) among all the candidate subjective sentences for our submission.

In particular, our three submissions are 1) Run2, answers derived from sentences retrieved by using emotional words of last year's task 2; 2) Run3, answers derived from sentences retrieved by using the emotional dictionary; and 3) Run1, the best (according to the probability score) 1000 among Run2 and Run3. In the following, due to the limited space, we will only show the results of our best run, Run1,

Table 2 Results of our System Submission 'RUN1'

Tasks	P	P@N	R	F	R-acc
Annotatior1					
Object Identification	**0.438**	**0.438**	**0.105**	**0.169**	**0.105**
Median	0.284	0.284	0.068	0.109	0.068
Max	0.438	0.438	0.105	0.169	0.105
Polarity Classification	**0.662**	**0.662**	**0.158**	**0.255**	**0.158**
Median	0.374	0.374	0.089	0.144	0.089
Max	0.662	0.662	0.158	0.255	0.158
Both	**0.340**	**0.340**	**0.081**	**0.131**	**0.081**
Median	0.232	0.232	0.055	0.089	0.055
Max	0.353	0.353	0.084	0.136	0.084
Annotator2					
Object Identification	0.348	0.348	0.087	0.139	0.087
Median	0.237	0.237	0.059	0.095	0.059
Max	0.354	0.354	0.089	0.142	0.089
Polarity Classification	0.612	0.612	0.153	0.245	0.153
Median	0.380	0.380	0.095	0.152	0.095
Max	0.612	0.612	0.153	0.245	0.153
Both	0.295	0.295	0.074	0.118	0.074
Median	0.198	0.198	0.050	0.079	0.050
Max	0.330	0.330	0.083	0.132	0.083
Annotator3					
Object Identification	0.321	0.321	0.089	0.139	0.089
Median	0.225	0.225	0.062	0.097	0.062
Max	0.321	0.321	0.089	0.139	0.089
Polarity Classification	0.544	0.544	0.150	0.235	0.150
Median	0.336	0.336	0.093	0.145	0.093
Max	0.544	0.544	0.150	0.235	0.150
Both	0.268	0.268	0.074	0.116	0.074
Median	0.190	0.190	0.052	0.082	0.052
Max	0.309	0.309	0.085	0.134	0.085

[4] They stand for person name, pronoun, and noun.

4 Task5

Task 5 focuses on two main aspects, one is Opinion Retrieval the other is Sentimental Polarity Analysis. We will present the detailed work in the following sections.

4.1 System Description

The whole system can be divided into three main parts: Relevance Computing, Subjective Computing and Polarity Computing, of which the first two consisting the Topic Retrieval module and on this basis we do the polarity computing to get the final result. As to the document relevance computing, we use Lemur[5] as our fundamental platform. Also we use HaiLiang Segmenter[6] to do segmentation work before indexing them. To improve the performance of word segment, we also manually add lots of words into the user defined word dictionary, which includes Sougou[7] popular word and the words which appear in the 50 given topics as well as their query expansive words. We adopt the classical BM25 language model to rank the relevance. As to the Subjective Computing, we adopt two methods to combine the topic relevance score and document opinionate score, one is just the simple linear way (Liu and Zhao, 2008) and the other is a complex quadratic method (Zhang and Ye, 2008), we will expatiate the details in the following section. As to Polarity Computing, we also adopt two models, one is based on the whole article, we adopt Polynomial Kernel model (Quan and Ren, 2008), which showed to be the best model in the last year's contest, the other is based on the extracted files, we compute by analyzing each sentence, considering some special Chinese sentence structures, to get the final polarity.

4.2 Query Expansion

Query Expansion is an important step in the whole processing. Based on last year’s test corpus, we summarize and follow the following rules:

1. Each topic should be expanded to its upper and lower level in the concept ontology tree, that is, we should consider its category and its content. For example, a film director should be expanded lower to the works he directed; also a film start should be expanded upper to the films he acted in.

2. Each topic as well as its expanded words should be existing in no more than 300 docs in the whole corpus. We do so mainly because too many related docs will cause too much noisy, we pursue precision more than recall.

3. For some unfamiliar and abstract topics like some economic items, we expand the topic manually by searching the topic in the test corpus and then pick the needed expansions.

---

[5] http://www.lumerproject.org

[6] http://www.hylanda.com/

[7] http://www.sogou.com/labs/

4.3 Opinion Retrieval

This will be divided into two parts: topic relevance scoring and subjective scoring. The Opinion Retrieval Score will be computed based on these two values. We mainly reference the method (Zhang and Ye) and the method (Liu and Zhao) of last year's contest. Due to space limitation, we only give the formulation used here, more theory and derivation details can be referenced in the references motioned above.

As to topic relevance scoring, we adopt the classical BM25 language model: given a document D and a query Q(including its expansion words), we want to compute score(D,Q) which implies the probability how q is related to d. The BM25 is shown as:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{\text{avgdl}})},$$

where $f(q_i,D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, and *avgdl* is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters, b is usually chosen as 0.75 and k1 values in the range [1.0,2.0]. IDF(q_i) is the IDF weight of the query term q_i. It is usually computed as:

$$\text{IDF}(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5},$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i

As to subjective score, we just compute the sentiment word frequency in the extracted doc. We express it as score(s|D):

score(S|D) = Sum(Count(s,D) / Count(D)) for all s $\in$ S-dict

Where Count(s,D) is the number of sentiment word s existing in the document D, Count(d) is the number of words in d and S-dict is the sentiment dictionary.

Finally we want to compute the topic retrieval score P(Q,S|D), we use both linear:

P(Q,S|D) = λ * score(D,Q) + (1 -λ) * score(S|D)

and quadratic combination:

P(Q,S|D) = score(D,Q) * score(S|D)

In our experiments, we take λ as 0.8 according to last year’s test corpus.

4.3 Sentiment Analysis

Here we have two methods for computing the sentimental polarity, we have experiments on both the whole doc and the extracted sentences, correspondingly we have two matched method, one is Polynomial Kernel model (Quan and Ren, 2008) and the other is based on polarity word analysis:

4.3.1 Polynomial Kernel

$$K(d_1,d_2) = (k(d_1,d_2)+c)^d \qquad (1)$$

Where $K(d_1,d_2)$ stands for the revised polynomial kernel value; $k(d_1,d_2)$ stands for the polynomial kernel value: $k(d_1,d_2)=\sum_{j=1}^{N} tf(t_j,d_1)tf(t_j,d_2)$ (2)

In which, $tf(t_i, d_j)$ stands for the frequency that word *i* appears in the document *j* The final equation to score the sentimental polarity is determined by sum(doc,Doc-pos) – sum(doc,Doc-neg), where, $sum(doc_i, Doc)$ is the sum of positive words and negative words in document i.

4.3.2 Scoring based on Word Polarity Analysis

This method is supposed that every sentiment word in the sentence must be selected and marked, and then calculate the amount of the sentiment word, in consideration of the sentences structure, we will get the scoring for the whole article. As we process on the concrete sentence, so we make full use of the context effect for sentiment polarity, for we have considered the modified word, the relational part and so on. We compute mainly based on our last year's method (Liu and Liu, 2008), More details can be got in the reference.

4.4 Experiments and Results

We altogether submit 9 runs; each of them is a combination of subjective score and sentiment polarity score, of which subjective score is also divided into two situations linear and quadratic. We rank them on last year's corpus and finally we get the following result:

Table 3 Result of Task 5

Run tag	MAP	P	R	F
SJTUCS-DFKILTTask5Run1	0.6254	0.124117	0.338402	0.162355
SJTUCS-DFKILTTask5Run2	0.5463	0.124072	0.338402	0.162284
SJTUCS-DFKILTTask5Run3	**0.6295**	0.119934	0.33988	0.158417
MEDIAN	0.5369	0.123912	0.338402	0.155235
MAX	**0.6298**	0.185867	0.396062	0.219588

As shown in the table, our MAP metric is good, almost close to the best result; other metrics are all beyond the average score, but not quite to our satisfaction， this is mainly because our combination processing of polarity score and subjective score may be not perfect, what's more, the pooling method may itself has some effect on the final result. As you may notice, the best score for the PRF and R-accuracy value is not quite high; we will have more improvement in the future.

5 Conclusion and Future Work

We are consistently devoting ourselves to Chinese NLP research work. In this evaluation contest we take part in three tasks and have a deeper understanding of Chinese processing. Although totally we get good results in this contest, we still have much space to improve, we also learn a lot from other participating organizations. After this contest, we will further devote us to the more detailed opinion analysis research work and get ready for the next COAE contest.

Acknowledgments

The second author is supported by the PIRE scholarship program; the fifth author thanks DFKI and the Cluster of Excellence on Multimodal Computing and Interaction for their support of the work; the other authors are financially supported by the National Science Foundation of China (Grant No. 60773087) and the UDS-SJTU Joint Research Lab for Language Technology.

References

Bing Liu.Opinion Mining and Summarization. WWW-2008, April 21, 2008 in Beijing, China.

Doran, C., Egedi, D., Hockey, B. A., Srinivas, B., Zaidel, M.XTAG System-A wide coverage grammar for English**.***COLING'94,* 1994

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of- Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

Xiwen FANG, Tianfang YAO. 2008. Design andImplementation of Chinese Polarity Dictionary in the First National KeenageWorkshop

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proc. of the 2nd International Conference on Knowledge Capture*, pages 70–77, Sanibel Island, Florida, USA.

Charlotta Engstrom. 2004. Topic dependence in sentiment classification. Master's thesis, University of Cambridge.

Eisner, J. 1996. Three new probabilistic models for dependency parsing: An exploration. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pages 340–345, Copenhagen, Denmark.

Hajic, J., Ciaramita, M., Johansson, R., Kawahara, D., Marti, M. A., Marquez, L., Meyers, A., Nivre, J., Pado, S., Stepanek, J., Stranak, P., Surdeanu, M., Xue, N., and Zhang, Y. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple lan- guages. In Proceedings of the 13th Conference on Computational Natural Language Learning. Boulder, CO, USA.

Jun Liu, Quansheng Liu, Mosha Chen, Hongyan Song, Gaohui Huang, Xiaojun Zhang, Tianfang Yao. 2008. Analysis of the Evaluation Results for our Tasks in COAE2008.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajic. 2005. Non-projective Dependency Parsing using Spanning Tree Algorithms. In Proceedings of HLT/EMNLP 2005.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. "A Conditional Random Field Word Segmenter." In Fourth SIGHAN Workshop on Chinese Language Processing. 2005.

Zhang, Y., Wang R., and Oepen, S. 2009. *Hybrid Multilingual Parsing with HPSG for SRL*. In Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2009), Boulder, CO, USA.

Kang Liu, Jun Zhao. 2008. NLPR-OR: a New Opinion Retrieval System in COAE2008

Min Zhang and Xingyao Ye. 2008 A Generative Model to Unity Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval. In Proceeding of SIGIR. 2008. Page 411-418.

Changqin Quan, Fuji Ren. 2008. Chinese Sentimental Classification Based on Polynomial Kernel Method in COAE2008.