Overview of Topic-based Chinese Message Polarity Classification in SIGHAN 2015

Xiangwen Liao
College of Mathematics and Computer Science, Fuzhou University, China
liaoxw@fzu.edu.cn

Binyang Li
School of Information Science and Technology, University of International Relations,
byli@uir.cn

Liheng Xu
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences,
lhxu@nlpr.ia.ac.cn

Abstract

This paper presents the overview of Topic-based Chinese Message Polarity Classification in SIGHAN 2015 bake-off. Topic-based message polarity classification plays an important role in sentiment analysis, information extraction, event tracking, and other related research areas. This task is designed to evaluate the techniques for Chinese message polarity classification towards a given topic. The task organizers manually constructed 25 topics together with 24,374 corresponding messages which were annotated to construct the training and testing datasets. The evaluation results achieved by the participators provide good suggestion for the future research.

1 Introduction

Recently, with the popularity of social media, such as microblogs, weblogs, and discussion forums, interests in analyzing sentiment and mining opinions in user-generated contents has grown rapidly. There are much work focusing on the overall polarity identification of a sentence, paragraph, or the document (Wiebe et al., 2005; Hu and Liu, 2004; Pang et al., 2002), without the consideration of the message polarity classification towards a specific topic. To this end, SIGHAN 2015 proposes a Topic-based Chinese Message Polarity Classification (TCMPC) task, which targets on classifying the polarity to the given topic in Chinese messages.

The task of Topic-based Chinese Message Polarity Classification is motivated by the need of microblog search where users attempt to discover popular sentiments on a topic. Similar pilot task has been proposed in the Chinese Opinion Analysis Evaluation (COAE) since 2008 (Zhao et al., 2008; Xu et al., 2009), which aimed at the document level based on blog corpus. Generally speaking, the mainstream techniques for COAE 2008 followed the thoughts of information retrieval, and adopted two-step approaches that first retrieved the documents relevant to the query, i.e. topic, and then identify the polarity for those retrieved documents. (Xu et al., 2009)

Currently, as the social media become popular, much research turned towards on short texts, e.g. messages. The task of Topic-based Chinese Message Polarity Classification in SIGHAN 8 bake-off is designed on the basis of task of Sentiment Analysis in Twitter in SemEval 2015 workshop. (Rosenthal et al., 2015) In this task, the organizers provide a collection of messages corresponding to a given topic and restricted sentiment resources which contain partial list of sentiment words. Participants are required to classify the topical messages into positive, negative, or neutral. This task is similar to COAE 2008 and 2009, but it focuses on sentiment polarity classification in short texts.

In the remainder of this paper, we first describe the task of topic-based message polarity classification. We then describe the process of data collection and annotation. We list and briefly describe the participating systems, and the results in our evaluation. Finally, we conclude and review the evaluation for future research.

2 Task Description

Topic-Based Chinese Message Polarity Classification is motivated by the function of microblog search where users attempt to discover popular sentiments towards on a topic.

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 56–60,
Beijing, China, July 30-31, 2015.

Organizers collect messages from Chinese microblog platforms[1] according to the predefined topics. Example 1 gives the sample of a topic together with the messages.

<Topic> "iphone6" (TopicID 0) **</Topic>**
<M15113801> 苹果公司已经发布了新产品 iphone6。 **</M15113801>**
<M15113803> iphone6 运行速度快，还是不错的。**</M15113803>**
<M15113805> 但是，iphone6 好像太薄了，容易折断，另外摄像头怎么是凸出的啊？？ **</M15113805>**

Example 1: Sample of input.

The participants are required to classify whether the message is of positive, negative, or neutral sentiment towards the given topic. For messages expressing both a positive and negative sentiment towards the topic, whichever is the stronger sentiment should be chosen. The analysis results are defined in the following format: <*runID; topicID*; *evalID*; *mesID*; *Polarity*>.

- *runID* is the team name of each participant;
- *topicID* is the name of each topic;
- *evalID* denotes different runs for the team;
- *mesID* is message *ID*;
- *Polarity* can be predicted sentiment polarity of topic (1 for positive, -1 for negative and 0 for neutral).

The first run by *team 1* of sample 1 is expected to be returned as follows:

<1; 0; 1; *M15113801*; 0>
<1; 0; 1; *M15113803*; 1>
<1; 0; 1; *M15113805*; -1>

In this task, the participants are required to submit two kinds of results based on: (1) restricted resource for fair comparison, e.g. sentiment lexicon, corpus; and (2) unrestricted resource. We believe that a freely available, annotated corpus that can be used as a common testbed is needed in order to promote research that will lead to a better understanding of how opinions are expressed in microblogs.

3 Datasets

In this section, we will describe our data collection and annotation.

3.1 Data Collection

We first identify the popular topics that widely arouse people's comments and sentiments from the newspapers. For this purpose, we utilized conventional topic detection techniques for detecting hot topics over a three months spanning from January 2015 to March 2015. Then, we also did some manual selection for the topics. First, we excluded topics that were incomprehensible, ambiguous, or were too general. Second, we removed microblogs that were just mentioning the topic, but not really about the topic, e.g. advertisements.

Given the set of identified topics, we further crawled the microblogs from the Chinese microblog platforms during the same time period that involved the topics. There were 24,374 messages among 25 topics in total, and the topics of test data were different from training data. In practice, most of the collected microblogs were likely to concentrate in the neutral class. To avoid class imbalance, we removed messages without sentiment-bearing words using NTUSD[2] as the repository of sentiment words.

3.2 Annotation

Three annotators were trained to annotate the dataset independently. Given a collection of messages, the annotation task is to label each message as positive, negative, or neutral with respect to the given topic. To avoid conflict, we pruned the messages which were classified into three categories by different annotators.

The Kappa coefficient indicating agreement was 0.8832 for the positive/negative classification and was 0.7829 for fine-grained annotation, where the annotator should annotate the stronger sentiment when both positive and negative sentiments towards the topic. Some statistics of the annotation results are displayed in Table 1 and Table 2. 538 out of 4,905 messages are labeled as negative accounting for 10.97%, while 394 messages are labeled as positive accounting for 8.03% in the training set. 3639 out of 19,469 messages are labeled as negative accounting for 18.69%, while 1152 messages are labeled as positive accounting for 5.91% in the testing set.

Table 1: Training dataset statistics.

Topics	Neg.	Neu.	Pos.	Total
三星 S6	95	646	246	987
疯抢日本马桶	168	776	29	973
央行降息	42	848	94	984
油价	108	880	9	997
雾霾	125	823	16	964
Total	538	3973	394	4905

[1] http://weibo.com
[2] http://www.datatang.com/data/44317/

Table 2: Testing dataset statistics.

Topics	Neg.	Neu.	Pos.	Total
12306 验证码	614	330	47	991
也门撤侨	4	951	42	997
何以笙箫默	33	852	115	1000
刘翔退役	28	817	137	982
跨省买墓	226	690	1	917
天使的城	5	951	39	995
孙楠退赛	142	828	13	983
少年四大名捕	17	940	40	997
就业季	392	540	4	936
延迟退休	438	522	27	987
换头手术	245	640	84	969
日修改教科书	333	630	4	967
日现大量中国游客	387	544	41	972
沪指 4000 点	29	844	103	976
漳州 PX 项目	48	945	2	995
美图手机	28	773	191	992
陶华碧	37	684	193	914
隆平超级稻	44	949	5	998
香港反水客	564	352	7	923
黄冈辉煌不再	25	896	57	978
Total	3639	14678	1152	19469

4 Evaluation Metrics

In the evaluation, both the resource-restricted and resource-unrestricted runs were adopted the same metrics. The messages were categorized into three classes, i.e., to assign one of the following three labels: positive, negative or objective/neutral. We evaluated the systems in terms of precision, recall, and F1 score for predicting positive and negative messages, respectively, Then we used macro-averaged F1 score for system comparison in the evaluation.

$$precision = \frac{System.Correct}{System.Proposed} \quad (1)$$

$$recall = \frac{System.Correct}{Golden} \quad (2)$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

$$Macro - F = \frac{F^{+} + F^{-}}{2} \quad (4)$$

5 Evaluation

Table 3 summarizes the submission statistics for 13 participant teams. Among 17 registered teams, 13 teams submitted their testing results of the Topic-based Chinese Message Polarity Classification. For this task, each participant is required to submit two kinds of results based on: restricted resource and unrestricted resource. Finally, we received 12 results based on restricted resource and 12 results based on unrestricted resource as shown in Table 3.

Table 4 showed the testing results based on restricted resource of the TCMPC task, and Table 5 showed the testing results based on unrestricted resource of the TCMPC task. In addition to *precision*, *recall* and *F1*, there are other fine-grained performance criteria, i.e., *precision+* reflects the percentage of correct positive messages among the positive messages submitted by each team; and *recall-* reflects the percentage of correct negative messages submitted by each team among the negative messages in dataset.

For general evaluations, the team TICS-dm achieved promising results in both restricted and unrestricted resources. Their results were about 10% higher than the second ranked team. Team ZWK, NEUDM1 and NEUDM2 also achieved nearly 75% performances. In general, most of teams perform better on unrestricted resource than restricted resource.

For fine-grained evaluations, the team TICSdm performed even more outstanding than other teams, i.e., their positive results were about 30% higher than the second ranked team on unrestricted resource. The team HLT HITSZ also performed well, i.e., their positive results were about 10% higher than the third ranked team on unrestricted resource. Overall, each team performed better on negative messages than positive messages.

6 Conclusion

This paper provides an overview of SIGHAN 2015 Bake-off Task 2: Topic-Based Chinese Message Polarity Classification, including task design, data preparation, evaluation metrics, and performance evaluation results. The task requires each participant to submit two kinds of result based on restricted resource for fair comparison and unrestricted resource. Regardless of actual performance, all submissions contribute to the common effort to produce an effective Chinese message polarity classifier, and the individual report in the bake-off proceedings provide useful insight into Chinese language processing. We believe that a freely available, annotated corpus that can be used as a common testbed is needed in order to promote research that will lead to a better understanding of how sentiment is conveyed in microblogs. All datasets with gold standards are publicly available for research purposes.

Table 3: Submission statistics for all participants.

Participant (Ordered by name of institution)		Restricted	Unrestricted
Team Name	**Institution**		
LCYS TEAM	Beijing Institute of Technology	1	1
yhz	East China Normal University	1	0
MSIIP THU0	Multimedia Signal and Intelligent Information Processing Laboratory, Tsinghua University	1	1
NUSTM	Nanjing University of Science and Technology	1	1
CUCSas	National Broadcast Media Language Resources Monitoring & Research Center, Communication University of China	1	1
KUASISLAB	National Kaohsiung University of Applied Sciences	0	1
NEUDM1	Northeastern University, China	1	1
NEUDM2	Northeastern University, China	1	1
neu sighan	Northeastern University, China	1	1
SIGSDS SCAU	South China Agricultural University	1	1
HLT HITSZ	Shenzhen Graduate School, Harbin Institute of Technology	1	1
TICS-dm	Tecent Intelligent Computing and Search Lab	1	1
ZWK	University of Montreal	1	1
Total		12	12

Table 4: Testing results based on restricted resource of the TCMPC task.

	Restricted						
	Pre.+	**Rec.+**	**F1+**	**Pre.-**	**Rec.-**	**F1-**	**Macro-F**
LCYS TEAM	0.2615	0.0590	0.0963	0.4023	0.1041	0.1655	0.1309
yhz	0.0364	0.0017	0.0033	0.2593	0.0879	0.1313	0.0673
MSIIP THU0	0.0988	0.0946	0.0967	0.3320	0.3768	0.3530	0.2249
NUSTM	0.1368	0.4922	0.2141	0.4052	0.5040	0.4492	0.3317
CUCSas	0.1202	0.2613	0.1647	0.3345	0.2336	0.2751	0.2199
NEUDM1	0.1418	0.1710	0.1551	0.3689	0.3528	0.3607	0.2579
NEUDM2	0.3188	0.0825	0.1310	0.4446	0.0827	0.1395	0.1353
neu sighan	0.0921	0.2977	0.1407	0.2700	0.1234	0.1694	0.1551
SIGSDS SCAU	0.1631	0.2813	0.2065	0.3607	0.3174	0.3377	0.2721
HLT HITSZ	0.2154	0.4045	0.2811	0.4584	0.6048	0.5216	0.4014
TICS-dm	0.6258	0.5139	0.5643	0.8232	0.4672	0.5961	**0.5802**
ZWK	0.2335	0.0920	0.1320	0.3047	0.1852	0.2304	0.1812

Table 5: Testing results based on unrestricted resource of the TCMPC task.

	Unrestricted						
	Pre.+	**Rec.+**	**F1+**	**Pre.-**	**Rec.-**	**F1-**	**Macro-F**
LCYS TEAM	0.1415	0.1128	0.1255	0.3635	0.1979	0.2562	0.1909
MSIIP THU0	0.1212	0.1788	0.1445	0.3412	0.3954	0.3663	0.2554
NUSTM	0.1767	0.5104	0.2626	0.4829	0.5191	0.5003	0.3815
CUCSas	0.1840	0.3602	0.2435	0.5011	0.3877	0.4372	0.3404
KUASISLAB	0.0886	0.0764	0.0821	0.2944	0.4089	0.3423	0.2122
NEUDM1	0.2696	0.1163	0.1625	0.4664	0.3333	0.3888	0.2757
NEUDM2	0.1763	0.0451	0.0719	0.4079	0.0566	0.0994	0.0857
neu sighan	0.0476	0.0564	0.0516	0.3296	0.3056	0.3171	0.1844
SIGSDS SCAU	0.1626	0.2899	0.2084	0.3784	0.3237	0.3489	0.2787
HLT HITSZ	0.2414	0.4167	0.3057	0.5159	0.5485	0.5317	0.4187
TICS-dm	0.5880	0.6207	0.6039	0.7918	0.6175	0.6938	**0.6489**
ZWK	0.1983	0.0200	0.0363	0.4072	0.0525	0.0930	0.0647

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 61300105), the Research Fund for Doctoral Program of Higher Education of China (No. 2012351410010), Fundamental Research Funds for the Central Universities (3262014T75, 3262015T20), the Key Project of Science and Technology of Fujian (No. 2013H6012), Shenzhen Fundamental Research Program (JCYJ20130401172046450) and the Project of Science and Technology of Fuzhou (No. 2012-G-113, 2013-PT-45). Special thanks to Hongfei Lin for providing the Chnese sentiment resources. We also thank Chen Chang, Yang Dingda, Ma Feixiang, Zhang liyao, Chen Xingjun for their annotation.

Reference

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pages 168-177, New York, NY, USA.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, pages 79-86, Philadephia, Pennsylvania, USA.

Sara Rosenthal, Preslav, Nakov, Svetlana Kiritchenko, Sif M Mohammad, Alan Ritter, Veselin Stoyanov. 2015. In Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval 2015.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2- 3):165-210.

Hongbo Xu, Tianfang Yao, Xuanjing Huang, Huifeng Tang, Feng Guan, and Jin Zhang. 2009. Overview of Chinese Opinion Analysis Evaluation 2009. In Proceedings of the Second Chinese Opinion Analysis Evaluation.

Jun Zhao, Hongbo Xu, Xuanjing Huang, Songbo Tan, Kang Liu, and Qi Zhang. 2008. Overview of Chinese Opinion Analysis Evaluation 2008. In Proceedings of the First Chinese Opinion Analysis Evaluation.