TeamX: A Sentiment Analyzer with Enhanced Lexicon Mapping and Weighting Scheme for Unbalanced Data

Yasuhide Miura
Fuji Xerox Co., Ltd. / Japan
yasuhide.miura@fujixerox.co.jp

Shigeyuki Sakaki
Fuji Xerox Co., Ltd. / Japan
sakaki.shigeyuki@fujixerox.co.jp

Keigo Hattori
Fuji Xerox Co., Ltd. / Japan
keigo.hattori@fujixerox.co.jp

Tomoko Ohkuma
Fuji Xerox Co., Ltd. / Japan
ohkuma.tomoko@fujixerox.co.jp

Abstract

This paper describes the system that has been used by TeamX in SemEval-2014 Task 9 Subtask B. The system is a sentiment analyzer based on a supervised text categorization approach designed with following two concepts. Firstly, since lexicon features were shown to be effective in SemEval-2013 Task 2, various lexicons and pre-processors for them are introduced to enhance lexical information. Secondly, since a distribution of sentiment on tweets is known to be unbalanced, an weighting scheme is introduced to bias an output of a machine learner. For the test run, the system was tuned towards Twitter texts and successfully achieved high scoring results on Twitter data, average F_1 70.96 on Twitter2014 and average F_1 56.50 on Twitter2014Sarcasm.

1 Introduction

The growth of social media has brought a rising interest to make natural language technologies that work with informal texts. Sentiment analysis is one such technology, and several workshops such as SemEval-2013 Task 2 (Nakov et al., 2013), CLEF 2013 RepLab 2013 (Amigó et al., 2013), and TASS 2013 (Villena-Román and García-Morera, 2013) have recently targeted tweets or cell phone messages as analysis text. This paper describes a system that has submitted a sentiment analysis result to Subtask B of SemEval-2014 Task9 (Rosenthal et al., 2014). SemEval-2014 Task9 is a rerun of SemEval-2013 Task 2 with different test data, and Subtask B is a task of message polarity classification.

The system we prepared is a sentiment analyzer based on a supervised text categorization approach. Various features and their extraction methods are integrated in the system following the works presented in SemEval-2013 Task 2. Additionally to these features, we assembled following notable functionalities to the system:

1. Processes to enhance word-to-lemma mapping.
   (a) A spelling corrector to normalize out-of-vocabulary words.
   (b) Two Part-of-Speech (POS) taggers to realize word-to-lemma mapping in two perspectives.
   (c) A word sense disambiguator to obtain word senses and their confidence scores.
2. An weighting scheme to bias an output of a machine learner.

Functionalities 1a to 1c are introduced to enhance information based on lexical knowledge, since features based on lexicons are shown to be effective in SemEval-2013 Task 2 (Mohammad et al., 2013). Functionality 2 is introduced to make the system adjustable to polarity unbalancedness known to exists in Twitter data (Nakov et al., 2013).

The accompanying sections of this papers are organized as follows. Section 2 describes resources such as labeled texts and lexicons used in our system. Section 3 explains the details of the system. Section 4 discusses the submission test run and some extra test runs that we performed after the test data release. Finally, section 5 concludes the paper.

2 Resources

2.1 Sentiment Labeled Data

The system is a constrained system, therefore only the sentiment labeled data distributed by the task

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 628–632,
Dublin, Ireland, August 23-24, 2014.

Type	#Used	#Full	%
Twitter(train)	6949	9684	71.8
Twitter(dev)	1066	1654	64.4
Twitter(dev-test)	3269	3813	85.7
SMS(dev-test)	2094	2094	100

Table 1: The numbers of messages for each type. 'train', 'dev', and 'dev-test' denote training, development, and development-test respectively. #Used is the number of messages that we were able to obtain, and #Full is the maximum number of messages that were provided.

Criterion	Lexicon
FORMAL	General Inquirer MPQA Subjectivity Lexicon SentiWordNet
INFORMAL	AFINN-111 Bing Liu's Opinion Lexicon NRC Hashtag Sentiment Lexicon Sentiment140 Lexicon

Table 2: The seven sentiment lexicons and their criteria.

organizers were used. However, due to accessibility changes in tweets, a subset of the training, the development, and the development-test data were used. Table 1 shows the numbers of messages for each type.

2.2 Sentiment Lexicons

The system includes seven sentiment lexicons namely: AFINN-111 (Nielsen, 2011), Bing Liu's Opinion Lexicon[1], General Inquirer (Stone et al., 1966), MPQA Subjectivity Lexicon (Wilson et al., 2005), NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013), Sentiment140 Lexicon (Mohammad et al., 2013), and SentiWordNet (Baccianella et al., 2010). We categorized these seven lexicons to two criteria: 'FORMAL' and 'INFORMAL'. Lexicons that include lemmas of erroneous words (e.g. misspelled words) were categorized to 'INFORMAL'. Table 2 illustrates the criteria of the seven lexicons. These criteria are used in the process of word-to-lemma mapping processes and will be explained in Section 3.1.3.

3 System Details

The system is a modularized system consisting of a variety of pre-processors, feature extractors, and a machine learner. Figure 1 illustrates the overview of the system.

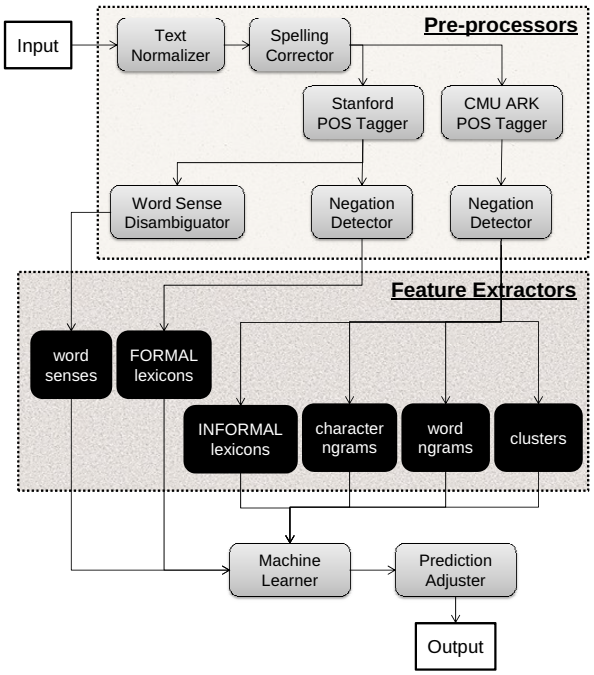

Figure 1: An overview of the system.

3.1 Pre-processors

3.1.1 Text Normalizer

The text normalizer performs following three rule-based normalization of an input text:

- Unicode normalization in form NFKC[2].
- All upper case letters are converted to lower case ones (ex. 'GooD' to 'good').
- URLs are exchanged with string 'URL's (ex. 'http://example.org' to 'URL').

3.1.2 Spelling Corrector

A spelling corrector is included in the system to normalize misspellings. We used Jazzy[3], an open source spell checker with US English dictionaries provided along with Jazzy. Jazzy combines DoubleMetaphone phonetic matching algorithm and a near-miss match algorithm based on Levenshtein distance to correct a misspelled word.

3.1.3 POS Taggers

The system includes two POS taggers to realize word-to-lemma mapping in two perspectives.

Stanford POS Tagger Stanford Log-linear Part-of-Speech Tagger (Toutanova et al., 2003) is one POS tagger which is used to map words

[1] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[2] http://www.unicode.org/reports/tr15/

[3] http://jazzy.sourceforge.net/

to lemmas of 'FORMAL' criterion lexicons, and to extract word sense features. A finite-state transducer based lemmatizer (Minnen et al., 2001) included in the POS tagger is used to obtain lemmas of tokenized words.

CMU ARK POS Tagger A POS tagger for tweets by CMU ARK group (Owoputi et al., 2013) is another POS tagger used to map words to lemmas of 'INFORMAL' criterion lexicons, and to extract ngram features and a cluster feature.

3.1.4 Word Sense Disambiguator

A word sense disambiguator is included in the system to determine a sense of a word. We used UKB[4] which implements graph-based word sense disambiguation based on Personalized PageRank algorithm (Agirre and Soroa, 2009) on a lexical knowledge base. As a lexical knowledge base, WordNet 3.0 (Fellbaum, 1998) included in the UKB package is used.

3.1.5 Negation Detector

The system includes a simple rule-based negation detector. The detector is an implementation of the algorithm on Christopher Potts' Sentiment Symposium Tutorial[5]. The algorithm is a simple algorithm that appends a negation suffix to words that appear within a negation scope surrounded by a negation key (ex. 'no') and a certain punctuation (ex. ':').

3.2 Features

The followings are the features used in the system.

word ngrams Contiguous 1, 2, 3, and 4 grams of words, and non-contiguous 3 and 4 grams of words are extracted from a given words. Non-contiguous ngram are ngrams where one of words are replaced with a wild card word '*'. Example of contiguous 3 grams is 'by_the_way', and the corresponding noncontiguous variation is 'by_*_way'.

character ngrams Contiguous 3, 4, and 5 grams of characters with in a word are extracted from given words.

lexicons Words are mapped to seven lexicons of section 2.2. For two sentiment labels (positive and negative) in each lexicon, following four values are extracted: total matched word count, total score, maximal score, and last word score[6]. For lexicons without sentiment scores, score 1.0 is used for all entries. Note that different POS taggers are used in word-to-lemma mapping as described in Section 3.1.3.

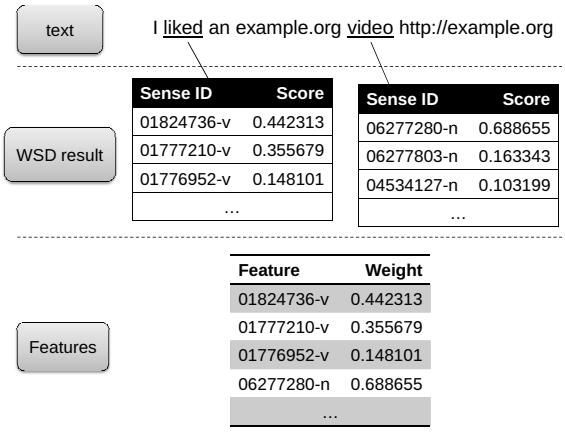

Figure 2: An example of word senses feature.

clusters Words are mapped to Twitter Word Clusters of CMU ARK group[7]. The largest clustering result consisting of 1000 clusters from approximately 56 million tweets is used as clusters.

word senses A result of the word sense disambiguator is extracted as weighted features according to their scores. Figure 2 shows an example of this feature.

The ngram features are introduced as basic bag-of-words features in a supervised text categorization approach. Lexicon features are designed to strengthen the lexical features of Mohammad et al. (2013) which have been shown to be effective in the last year's task. Cluster features are implemented as an improvement for an supervised NLP system following the work of Turian et al. (2010). Word sense features are utilized to help subjectivity analysis and contextual polarity analysis (Akkaya et al., 2009).

3.3 Machine Learner

Logistic Regression is utilized as an algorithm of a supervised machine learning method. As an implementation of Logistic Regression, LIBLINEAR (Fan et al., 2008) is used. A Logistic Regression is trained using the features of Section 3.2 with the three polarities (positive, negative, and neutral) as labels.

[4] http://ixa2.si.ehu.es/ukb/

[5] http://sentiment.christopherpotts.net/lingstruc.html#negation

[6] The total number of lexical features is $7 \times 2 \times 4 = 56$.

[7] http://www.ark.cs.cmu.edu/TweetNLP/

Parameter Selection Source	Parameters			Sources				
	C	w_{pos}	w_{neg}	LiveJournal 2014	SMS 2013	Twitter 2013	Twitter 2014	Twitter2014 Sarcasm
Twitter(train)+Twitter(dev)	0.07	1.7	2.6	71.23	62.33	71.28	70.40	53.32
Twitter(dev-test)*	0.03	2.4	3.3	69.44	57.36	72.12	70.96	56.50
SMS(dev-test)	0.80	1.1	1.2	72.99	68.92	65.65	66.66	48.24
SMS(dev-test)+Twitter(dev-test)	0.07	1.9	2.0	72.54	65.44	70.41	69.80	51.09

Table 3: The scores for each source in the test runs. The run with asterisk (*) denotes the submission run. The values in the 'Sources' columns represent scores in SemEval-2014 Task 9 metric (the average of positive F_1 and negative F_1).

3.4 Prediction Adjuster

Since the labels in the tweets data are unbalanced (Nakov et al., 2013), we prepared a prediction adjuster for Logistic Regression output. For each polarity l, an weighting factor w_l that adjusts a probability output $Pr(l)$ is introduced. An updated prediction label is decided by selecting an l that maximizes $score(l)$ which can be expressed as equation 1.

$$\arg\max_{l \in \{pos,neg,neu\}} \quad score(l) = w_l Pr(l) \qquad (1)$$

The approach we took in this prediction adjuster is a simple approach to bias an output of Logistic Regression, but may not be a typical approach to handle unbalanced data. For instance, LIBLINEAR includes the weighting option '-wi' which enables a use of different cost parameter C for different classes. One advantage of our approach is that the change in w_l does not require a training of Logistic Regression. Various values of w_l can be tested with very low computational cost, which is helpful in a situation like SemEval tasks where the time for development is limited.

4 Test Runs

4.1 Submission Test Run

The system was trained using the 8,015 tweets included in Twitter(train) and Twitter(dev) described in Section 2.1. Three parameters: cost parameter C of Logistic Regression, weight w_{pos} of the prediction adjuster, and weight w_{neg} of the prediction adjuster, were considered in the submission test run. For the w_{neu} of the prediction adjuster, a fixed value of 1.0 was used.

Prior to the submission test run, the following two steps were performed to select a parameter combination for the submission run.

Step 1 The system with all combinations of C in range of {0.01 to 0.09 by step 0.01, 0.1 to 0.9 by step 0.1, 1 to 10 by step 1}, w_{pos} in range of {1.0 to 5.0 by step 0.1}, and w_{neg} in range of {1.0 to 5.0 by step 0.1} were prepared[8].

Step 2 The performances of the system for all these parameter combinations were calculated using Twitter(dev-test) described in Section 2.1.

As a result, the parameter combination $C = 0.03$, $w_{pos} = 2.4$, and $w_{neg} = 3.3$ which performed best in Twitter(dev-test) was selected as a parameter combination for the submission run.

Finally, the system with the selected parameters was applied to the test set of SemEval-2014 Task 9. 'Twitter(dev-test)' in Table 3 shows the values of this submission run. The system achieved high performances on Twitter data: 72.12, 70.96, and 56.50 on Twitter2013, Twitter2014, and Twitter2014Sarcasm respectively.

4.2 Post-Submission Test Runs

The system performed quite well on Twitter data but not so well on other data on the submission run. After the release of the gold data of the 2014 test tun, we conducted several test runs using different parameter combinations. 'Twitter(train)+Twitter(dev)', 'SMS(dev-test)', and 'SMS(dev-test)+Twitter(dev-test)' are the results of test runs with different data sources used for the parameter selection process. In 'Twitter(train)+Twitter(dev)', the parameter combination that maximizes a micro-average score of 5-fold cross validation was chosen since the training data and the parameter selection are equivalent.

The parameter combination selected with 'Twitter(train)+Twitter(dev)' showed similar result as the submission run, which is high performances on Twitter data. In the case of 'SMS(dev-test)', the system performed well on 'LiveJournal2014' and 'SMS(dev-test)' namely 72.99 and 68.92. How-

[8]The total number of parameter combination is $29 \times 51 \times 51 = 75429$.

ever, in this parameter combination the scores on Twitter data were clearly lower than the submission run. Finally, 'SMS(dev-test)+Twitter(dev-test)' resulted to a mid performing result, where scores for each source marked in-between values of 'Twitter(dev-test)' and 'SMS(dev-test)'.

5 Conclusion

We proposed a system that is designed to enhance information based on lexical knowledge and to be adjustable to unbalanced training data. With parameters tuned towards Twitter data, the system successfully achieved high scoring results on Twitter data, average F_1 70.96 on Twitter2014 and average F_1 56.50 on Twitter2014Sarcasm.

Additional test runs with different parameter combination showed that the system can be tuned to perform well on non-Twitter data such as blogs or short messages. However, the limitation of our approach to directly weight a machine learner's output was shown, since we could not find a general purpose parameter combination that can achieve high scores on any types of data.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments to improve this paper.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of EACL 2009*, pages 33–41.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of EMNLP 2009*, pages 190–199.

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2013. Overview of RepLab 2013: Evaluating online reputation monitoring systems. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*, pages 2200–2204.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. In *Journal of Machine Learning Research*, volume 9, pages 1871–1874.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, pages 321–327.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, pages 312–320.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL 2013*, pages 380–390.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the eighth international workshop on Semantic Evaluation Exercises (SemEval-2014)*.

Philip Stone, Dexter Dunphy, Marshall Smith, and Daniel Ogilvie. 1966. *General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

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*, pages 252–259.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

Julio Villena-Román and Janine García-Morera. 2013. TASS 2013 - Workshop on sentiment analysis at SEPLN 2013: An overview. In *Proceedings of the TASS workshop at SEPLN 2013*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP 2005*, pages 347–354.