UWB Participation in the Multiling's OnForumS Task

Peter Krejzl, Josef Steinberger
Dept. of Computer Science and Eng.,
Faculty of Applied Sciences,
University of West Bohemia,
Univerzitní 8, 306 14 Plzeň
Czech Republic
pkrejzl@gmail.com
jstein@kiv.zcu.cz

Tomáš Hercig, Tomáš Brychcín
NTIS – New Tech. for the Inf. Society,
Faculty of Applied Sciences,
University of West Bohemia,
Univerzitní 8, 306 14 Plzeň
Czech Republic
tigi@kiv.zcu.cz,
brychcin@kiv.zcu.cz

Abstract

This paper presents a system used for the Online Forum Summarization task of Multiling 2015. We drafted an approach to all 3 subtasks: linking comment sentences to relevant content of the article, detecting sentiment polarity of the comment and agreement between the linked texts. For the comment linking we use vector space model and latent dirichlet allocation. The sentiment and argument structure labeling is based on a maximum entropy classifier. The preliminary results indicate a good precision for English but worse for Italian.

1 Introduction

The Multiling shared tasks were traditionally linked to summarization of news articles (Giannakopoulos, 2013; Giannakopoulos et al., 2011). However, the increasing amounts of user-supplied comments in most major online news portals suggest the need for automatic summarization methods, which brings a novel challenge for the summarization community. The comments are related either to specific points within the article or to previous comments.

The new edition of Multiling shared tasks (2015) brings a new task: Online Forum Summarization (OnForumS). The purpose of the OnForumS track is to set the ground for investigating how such a mass of comments can be summarized. An important initial step in developing reader comment summarization systems is to determine what comments relate to, be that either specific points within the text of the article, the global topic of the article, or comments made by other users. This constitutes a linking task. Furthermore, a set of link types or labels may be articulated to capture whether, for example, a comment agrees with, elaborates, disagrees with, etc., the point made in the commented-upon text. Solving this labelled linking problem should facilitate the creation of reader comment summaries by allowing, for example, that comments relating to the same article content can be clustered, points attracting the most comment can be identified, representative comments can be chosen for each key point, and the implications of labelled links can be digested (e.g., numbers for or against a particular point), etc.

The OnForumS task is a particular specification of the linking task, in which systems take as input a news article with a reduced set of comments (sifted, according to predefined criteria, from what could otherwise be thousands of comments) and are asked to link and label each comment to sentences in the article (which, for simplification, are assumed to be the appropriate units here) or to preceding comments. The labels include agreement/disagreement and sentiment indicators. The data cover two languages (English and Italian).

We drafted our first approach to resolve the issues. This paper describes our system which participated in the OnForumS task. We first discus two technologies behind our run: semantic analysis and sentiment analysis (section 2). Then, we move to the description of the system (section 3) and discussion of preliminary results of system's precision (section 4).

2 Technologies behind

2.1 Semantic analysis

The backbone principle of methods for discovering hidden meaning in a plain text is the formulation of the *Distributional Hypothesis* in (Firth, 1957): "a word is characterized by the company it keeps." The direct implication of this hypothesis is that the meaning of a word is related to the context where it usually occurs and thus it is possible to compare the meanings of two words by

statistical comparisons of their contexts. This implication was confirmed by empirical tests carried out on human groups in (Rubenstein and Goodenough, 1965; Charles, 2000).

To represent the meaning at the document level, the *Bag-of-words Hypothesis* was shown to be useful. The term *bag* means a *set* where the order has no role, however, the duplicates are allowed (the bags a, a, a, b, b, c and c, a, b, a, b, a are equivalent).

The first practical application of the hypothesis was arguably in information retrieval. In work of (Salton et al., 1975), the documents were represented as bags-of-words and the frequencies of words in a document indicated the relevance of the document to a query. The implication is that two documents tend to be similar if they have similar distribution of similar words, no matter what is their order. This is supported by the intuition that the topic of a document will probabilistically influence the author's choice of words when writing the document.

Similarly, the words can be found related in meaning if they occur in similar documents (where document represents the word context). Thus, both hypotheses (bag-of-words hypothesis and distributional hypothesis) are related.

The models based on the Distributional Hypothesis or the Bag-of-words Hypothesis typically represent the meaning as a vector. Represented geometrically, the meaning is a point in a high-dimensional space. The documents that are closely related in meaning tend to be closer in the space. As a measure of the similarity between two documents, we use the cosine similarity calculated as the cosine of the angle between the corresponding vectors.

2.1.1 Vector Space Model (VSM)

We build the co-occurrence matrix $\mathbb{M} = |D| \times |W|$, where $|D|$ is the size of document collection and $|W|$ is the size of the word vocabulary. Each element of the matrix corresponds to the count of how many times the word $w \in W$ occur in the document $d \in D$. These elements are then weighted according to TF-IDF (term frequency – inverse document frequency) scheme (Manning et al., 2008).

The meaning of the document d is represented as an appropriate row vector (TF-IDF values of words in document d).

2.1.2 Latent Dirichlet Allocation (LDA)

LDA(Blei et al., 2003) is a well known topic model. LDA is based on the Distributional Hypothesis and the Bag-of-words Hypothesis, i.e., that the word order does not matter and there is some latent relation between the words within the same document (within the same context).

LDA assumes the documents are mixtures of topics and each topic is assumed to be a mixture of words. We use Gibbs sampling to infer the topic assignments (Griffiths and Steyvers, 2004).

We represent the meaning of a document d as a K-dimensional vector, where K is the number of topics in LDA. Each value in this vector is set to be the probability of the topic z conditioned on the corresponding document d, where $1 \leq z \leq K$. LDA assumes these probabilities are drawn from Dirichlet distribution.

We use the LDA implementation from the MALLET (McCallum, 2002) software package. The hyperparameters of the Dirichlet distributions were initially set to $\alpha = 50/K$ (for the distribution of topics in a document) and $\beta = 0.1$ (for the distribution of words in a topic). This setting is recommended by (Griffiths and Steyvers, 2004).

We use 100 topics LDA ($K = 100$) in our experiments.

2.2 Sentiment analysis

We trained the Maximum Entropy (MaxEnt) classifier on out of domain data sets. The training was done using a Java framework for machine learning *Brainy* (Konkol, 2014). IT training dataset comes from the Sentipolc 2014 (Basile et al., 2014) development data. We were able to retrieve only 3420 documents from the original 4513 documents. The EN dataset consists of the facebook dataset in (Zhang et al., 2011) and IMDb dataset in (Pang et al., 2002). Table 1 contains statistics of the used datasets.

Dataset	positive	neutral	negative	total
IT	827	1161	1432	3420
EN	1641	280	1079	3000

Table 1: Sentiment label distribution in training datasets.

2.2.1 Features

For the model training we used the following language-independent features.

Character n-gram We used character n-gram features (Blamey et al., 2012). We set the minimum occurrence of a particular character n-gram to 5. Our character feature set contains 3-grams to 6-grams.

N-gram We used word unigrams, bigrams and trigrams as binary features. The feature space is pruned by the minimum n-gram occurrence set to five.

Skip-bigram Instead of using sequences of adjacent words (n-grams) we used skip-grams (Guthrie et al., 2006), which skip over arbitrary gaps. We consider skip-bigrams with two to five word skips and remove skip-grams with a frequency ≤ 20.

Emoticons We used two lists of positive and negative emoticons (Montejo-Ráez et al., 2012). The feature captures the number of occurrences of each class of emoticons within the text.

3 The system description

The system processes all comment sentences and calculates their similarities to article sentences or parent comment sentences. The "similarity" score is based on the two models discussed in section 2.1: VSM and LDA. The final score is calculated as an average of similarities computed using both the models. At the end of this phase, there is a list of link candidates: either comment sentence to article sentence or comment sentence and comment sentence. Candidates with the anchor sentence shorter than six words are filtered out. The final output of our system consists of one percent of links. The system selects those with the largest similarity score. For training the VSM and LDA models we used the TREC data. In particular, for English it was Glassgow Herald 1995 and Los Angeles Times 1994 and 2002 datasets.

The next step was to calculate sentiment polarities. For each detected link, sentiments of both sentences were calculated. It was classified into three classes: positive, neutral and negative. The sentiment of the comment was used to fill the sentiment label of the task. Both the comment sentence and the linked sentence sentiments were used to assign the agreement (argument structure) label. Table 2 describes the simplest method to derive the label in the in-favour/against/impartial scale.

4 Results

The links identified by the system went through validation in Crowd Flower. The contributors were asked to judge whether the two shown sentences are related. In the case of the 'yes' answer they validated also the detected sentiment and the argument structure.

The test set contains 10 English articles and 5 Italian. 9 systems participated for English and 7 for Italian. At this moment, we have only results of precision. Table 3, resp. table 4, shows precision and rank of our system for English, resp. Italian.

run	linking	argument	sentiment
best	.928	.990	.946
UWB	**.851 (4)**	**.974 (3)**	**.897 (5)**
average	.829	.896	.897
worst	.702	.859	.874

Table 3: OnForumS results - English: precision (rank). 9 systems participated in total.

In 5 of the 10 English articles, all the links proposed by our system were correct. It was ranked 4th (out of 9). All prediction of argument structure were correct in 8 articles. Our run was ranked 3rd with a very large precision (.974). In 7 articles, all sentiment predictions were correct, ranking our system 5th, also with a very large precision (.897).

run	linking	argument	sentiment
best	.590	1	.666
UWB	**.250 (2)**	**1 (1)**	**.250 (5)**
average	.152	0.750	.333
worst	.010	0	0

Table 4: OnForumS results - Italian: precision (rank). 7 systems participated in total.

Results for Italian seem to be opposite. In 2 of the 5 Italian articles, all the links proposed by our system were wrong, although our system was ranked 2nd. All prediction of argument structure were judged as correct ranking our system at the top of the list. On the other side is sentiment, for which only 1/4 of the sentiment labels was positively validated.

5 Conclusion

We took part in the summarization community effort to initiate research towards summarizing comments related to an article in multiple languages.

	linked: POSITIVE	linked: NEUTRAL	linked: NEGATIVE
comment: POSITIVE	IN FAVOUR	IN FAVOUR	AGAINST
comment: NEUTRAL	IMPARTIAL	IMPARTIAL	IMPARTIAL
comment: NEGATIVE	AGAINST	AGAINST	IN FAVOUR

Table 2: Comparing the comment sentence and the linked sentence polarities to derive the argument label.

The pilot of the OnForumS task included English and Italian data and aimed at the first step of the summarization process: linking comments to article sentences and labeling their sentiment and argument structure. We proposed an approach based on semantic and sentiment analysis. Based on the preliminary results we see a real need to run the evaluation in multiple languages as the system performs surprisingly precisely for English but much worse for Italian. The most difficult subtask – deciding the agreement between two text – seems to work better than expected in both the languages.

Acknowledgments

This work was partly supported by project MediaGist, EU's FP7 People Programme (Marie Curie Actions), no. 630786.

References

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task. *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*.

Ben Blamey, Tom Crick, and Giles Oatley. 2012. R U : -) or : -(? character- vs. word-gram feature selection for sentiment classification of OSN corpora. In *Proceedings of AI-2012, The Thirty-second SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 207–212. Springer.

D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

W. G. Charles. 2000. Contextual correlates of meaning. *Applied Psycholinguistics*, 21(04):505–524.

John R. Firth. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32.

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2011. Tac 2011 multiling pilot overview. In *Proceedings of the Text Analysis Conference 2011*, Gaithersburgh, USA, November. NIST.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria, August. Association for Computational Linguistics.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4.

Michal Konkol. 2014. Brainy: A machine learning library. In Leszek Rutkowski, Marcin Korytkowski, Rafa Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing*, volume 8468 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Scoring, term weighting, and the vector space model*. Cambridge University Press. Cambridge Books Online.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña López. 2012. Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 3–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo, Ankit Agrawal, Diana Palsetia, Kathy Lee, Wei-keng Liao, and Alok Choudhary. 2011. Ses: Sentiment elicitation system for social media data. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 129–136. IEEE.