Unsupervised Negation Focus Identification with Word-Topic Graph Model

Bowei Zou **Qiaoming Zhu** **Guodong Zhou**[*]

Natural Language Processing Lab, School of Computer Science and Technology
Soochow University, Suzhou, 215006, China
zoubowei@gmail.com, {qmzhu, gdzhou}@suda.edu.cn

Abstract

Due to the commonality in natural language, negation focus plays a critical role in deep understanding of context. However, existing studies for negation focus identification major on supervised learning which is time-consuming and expensive due to manual preparation of annotated corpus. To address this problem, we propose an unsupervised word-topic graph model to represent and measure the focus candidates from both lexical and topic perspectives. Moreover, we propose a document-sensitive biased PageRank algorithm to optimize the ranking scores of focus candidates. Evaluation on the *SEM 2012 shared task corpus shows that our proposed method outperforms the state of the art on negation focus identification.

1 Introduction

Negation is used to reverse the polarity of part of statements that are otherwise affirmative by default (Blanco and Moldovan, 2011), which is common in natural language. *Negation focus* is defined as the special part in sentence, which is most prominently or explicitly negated by a negative expression. For example, sentence (1) could be interpreted as *He stopped, but not until he got to Jackson Hole* with a positive part *he stopped* and a negative part *until he got to Jackson Hole*.

(1) *He did**n't** stop until he got to Jackson Hole.*

Our previous work (Zou et al., 2014) showed that contextual information plays a critical role on negation focus identification. For better illustration of this conclusion, they manually analyze the evidences for 100 negation focuses. It is surprising that 77 focuses can be identified with help of contextual information. This indicates that negation focus is often related with what authors repeatedly states in context. In this paper, we thus focus on graph-based ranking methods (Mihalcea and Tarau, 2004) which first build a word graph according to word co-occurrences within document, and then use random walk algorithms (e.g., PageRank) to measure word importance.

However, for negation focus identification, the graph-based methods may suffer from the following two problems: (a) the words in graph-based methods are strongly connected by co-occurrence rather than semantic content, which do not necessarily guarantee that they are relevant to the negation focus in context; and (b) identifying a negation focus may be affected by not only the relatedness of surrounding words but also its importance in current document which is not considered in standard random walk algorithms.

To address the above problems, we propose a word-topic graph model by adding a topical layer on the original word layer to capture the semantic clues from both lexical and topic perspectives. Besides, a document-sensitive PageRank algorithm is also proposed to optimize the graph model by considering the document's major topics. Experimental results indicate that our word-topic graph model outperforms other baseline methods. Moreover, our model is unsupervised and requires only un-annotated text for training.

The rest of this paper is organized as follows. Section 2 overviews the related work. Section 3 introduces our word-topic graph model with contextual discourse information. Section 4 reports the experimental results and analysis. Finally, we conclude our work in Section 5.

[*] Corresponding author

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1655–1659,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

2 Related Work

So far there is little work on negation focus identification, which was pioneered by Blanco and Moldovan (2011) who investigated the negation phenomenon in semantic relations and proposed a supervised learning approach to identify the focus of a negation expression. However, although Morante and Blanco (2012) proposed negation focus identification as one of the *SEM'2012 shared tasks, only one team (Rosenberg and Bergler, 2012) participated in. They identified negation focus by three heuristic rules.

Our previous work (Zou et al., 2014) demonstrates the effectiveness of contextual information for negation focus identification. On this basis, we further optimize the graph model in both the topical layer and the PageRank algorithm in this paper.

In recent years, many algorithms are widely used to incorporate word graph models and topical information within random walk. Our work is originally inspired by Liu et al. (2010). Their method runs decomposed Topical PageRank (TPR) for each topic separately, and then calculates the word scores with respect to different topics. When setting the edge weights, only word co-occurrence is considered. Different from their work, our word-topic graph model runs on a two-layers (word layer and topical layer) graph model and sets the edge weights by measuring both word similarity and topic distribution.

3 Methods

The word-topic graph model consists of word layer and topical layer, as shown in Figure 1. While the word layer is constructed according to word co-occurrence within a sliding window, which expresses the cohesion relationship between words in the context, the topical layer is to refine the graph model over the discourse contextual information.

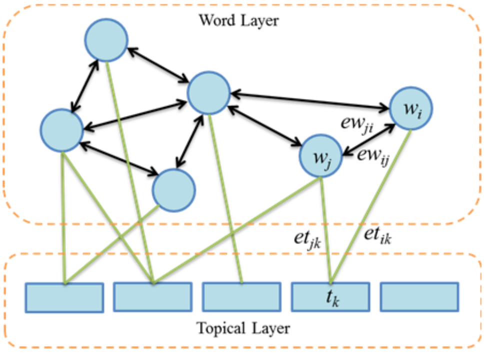

Figure 1. Word-topic graph model.

3.1 Constructing Word Layer

The word layer is constructed according to word co-occurrence within a sliding window, which expresses the cohesion relationship between words in the context. It can be denoted as L_{word} (W, E_w), where vertex set $W=\{w_i\}$ represents the set of words in one document and link set $E_w = \{e_{ij}|w_i, w_j \in W\}$ represents the set of directed edges between these words. Note that only content words are considered. Namely, we consider nouns, verbs, adjectives, and adverbs.

The link directions are added from the first word pointing to other words within a sliding s-width sentence window. Directed edge ew_{ij} is weighted to represent the relatedness between word w_i and word w_j in a document with transition probability $P_w(j|i)$ from i to j, which is normalized as follows:

$$P_w(j\,|\,i) = \frac{sim(w_i, w_j)}{\sum_{k:w_i \to w_k} sim(w_i, w_k)}, \tag{1}$$

where the denominator represents the out-degree of vertex w_i , and $sim(w_i,w_j)$ denotes the similarity between word w_i and w_j. In this paper, both corpus-based and knowledge-based methods are evaluated to calculate the similarity between words.

- Word co-occurrence. If word w_i and word w_j occur in a s-width sliding sentence window, $sim(w_i,w_j)$ increases 1.
- WordNet similarity (Miller, 1995). In this paper, we adopt the *path similarity* function to measure relatedness of nouns and verbs, and adopt the *similar to* function to measure relatedness of adjectives and adverbs by using the NLTK toolkit[1] (Bird et al., 2009).

Note that $sim(w_i,w_i) = 0$ to avoid self-transition, and $sim(w_i,w_j)$ and $sim(w_j,w_i)$ may not be equal.

3.2 Preliminaries for Topical Layer

To infer the latent topic distributions of words, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a typical of topic model, is directly applied. By the set of topics which derive from a corpus, we can obtain:

- $P(t|w)$, the probability of topic t given word w, which indicates how much that word w focuses on topic t, and
- $P(t|d)$, the probability of topic t given document d, which indicates how much that document d focuses on topic t.

[1] http://www.nltk.org/

Then, the similarity between two words or between word w_i and document d can be measured by the similarity between their corresponding topic distributions. Formally, we denote a topic distribution as θ, and measure the similarity by using:

- Dot-product. We consider the topic distributions as vectors and apply the dot-product, a geometrically motivated function, to calculate the similarity:

$$Inner(\theta_{w_i},\theta_{w_j}) = \theta_{w_i}\cdot\theta_{w_j} = \sum\nolimits_{t_k\in T} P(t_k \mid w_i)\cdot P(t_k \mid w_j), \quad (2)$$

$$Inner(\theta_{w_i},\theta_d) = \theta_{w_i}\cdot\theta_d = \sum\nolimits_{t_k\in T} P(t_k \mid w_i)\cdot P(t_k \mid d). \quad (3)$$

- Kullback Leibler (KL) divergence (Lin, 1991). Considering the asymmetry (Eq.(4)), we obtain a symmetrized measure by Eq.(5).

$$D(\theta_i,\theta_j) = \sum\nolimits_{t_k\in T} P_i(t_k)\log_2\frac{P_i(t_k)}{P_j(t_k)}. \quad (4)$$

$$D_{KL}(\theta_i,\theta_j) = D(\theta_i,\theta_j) + D(\theta_j,\theta_i). \quad (5)$$

Note that $D_{KL}(\theta_i, \theta_j)$ is undefined if $P_i(t_k)$=0 or $P_j(t_k)$=0 for any $t_k \in T$. For this reason, only the topics which make both $P_i(t_k)\neq 0$ and $P_j(t_k)\neq 0$ are adopted to calculate the KL divergence between two topic distributions.

3.3 Word-Topic Graph Model

The word layer can well capture the relatedness between words, but just partially model the negation focus since it is more directly related with topic than content. Therefore, we add one more layer to refine the graph model over the topical information, as shown in Figure 1. Formally, the word-topic graph is defined as $G_{topic}(W, T, E_w, E_t)$, where vertex set $T=\{t_i\}$ represents the set of topics in all of documents in corpus and link set $E_t = \{et_{ij}|w_i\in W,\ t_j \in T\}$ represents the set of undirected edges between words and topics.

Considering that the topical layer can provide more contextual semantic information, we refine the relatedness between words by using a topical transition probability $P_t(j|i)$ which is calculated by two kinds of measurements:

$$P_t(j \mid i) = \frac{sim(\theta_{w_i},\theta_{w_j})}{\sum\limits_{k:w_i\to w_k} sim(\theta_{w_i},\theta_{w_k})}. \quad (6)$$

Here, the similarity is measured by the dot-product or the KL divergence (using reciprocals). On this basis, the word transition probability $P_w(j|i)$ is updated as following:

$$P_w^{'}(j \mid i) = \mu\cdot P_w(j \mid i) + (1-\mu)\cdot P_t(j \mid i). \quad (7)$$

where $\mu\in[0,1]$ is the coefficient controlling the relative contributions from the lexical information and the topical information.

Moreover, the weights of word vertices are calculated by a PageRank algorithm. In standard PageRank (Page et al., 1998), words are set to be the same value, which indicates there is equal importance to all of words in a document. However, intuitively, we should allocate higher weights to those words with high relevance to the document. Therefore, we assign a document-sensitive value to word w_i:

$$R_d(w_i) = \frac{sim(\theta_{w_i},\theta_d)}{\sum\nolimits_{w_k\in W} sim(\theta_{w_k},\theta_d)} \quad (8)$$

and calculate the weights of word vertices iteratively by using a biased PageRank algorithm:

$$Score^{(n+1)}(w_i) = \lambda\sum\nolimits_{j\neq i} Score^{(n)}(w_j)P_w^{'}(j \mid i) + (1-\lambda)R_d(w_i). \quad (9)$$

All of the PageRank algorithms are terminated when the number of iterations reaches 100 or the difference of each vertex between consecutive iterations is less than 0.001.

Finally, according to the annotation guidelines (Blanco and Moldovan, 2011), the focus is always a full text of a semantic role. Thus, we select all of semantic roles in sentence as candidate focuses for ranking. The ranking score of a candidate focus f is computed by averaging the scores of all words inside the candidate:

$$score_{avg}(f) = \frac{\sum\nolimits_{w_i\in f} score(w_i)}{Count(f,\bullet)}, \quad (10)$$

where $count(f,\bullet)$ denotes the number of content words within the candidate. Then the top ranked candidate is chosen as the negation focus.

4 Experimental Results

To evaluate the performance of our word-topic graph model for negation focus identification, we carry out experiments on the *SEM'2012 shared task corpus[2]. As a freely downloadable resource, the corpus is annotated on top of PropBank, which uses the WSJ section of the Penn TreeBank. In particular, negation focus annotation on this corpus is restricted to verbal negations (Blanco and Moldovan, 2011). In total, this corpus provides 3,544 instances of negation focus annotations. Although for each instance, the corpus only provides the current sentence, the previous and next sentences as its context, we sort to

[2] http://www.clips.ua.ac.be/sem2012-st-neg/

the Penn TreeBank[3] to obtain the corresponding document as its discourse context. For fair comparison, we adopt the same partition as *SEM'2012 shared task in our experiments.

We evaluate our results in terms of accuracy. To see whether an improvement is statistically significant, we conduct significance testing using the paired t-test.

For estimating the topical transition probability $P_t(j|i)$ and the document-sensitive value $R_d(w_i)$, we employ GibbsLDA++[4], an LDA model using Gibbs Sampling technique for parameter estimation and inference (Griffiths, 2002). We set the parameters $\alpha = 50/T$ and $\beta = 0.1$ as Griffiths and Steyvers (2004) suggested.

4.1 Influences of Parameters

There are two major parameters in our models that may influence the performance, including: (a) the damping factor μ of the word transition probability $P'_w(j|i)$ (Eq.(7)) and (b) the damping factor λ of the word-topic graph model (Eq.(9)).

Figure 2 shows the accuracy when varying μ from 0.1 to 0.9 with an interval of 0.1 and when varying λ from 0.05 to 1 with an interval of 0.05. We notice that the best performance is achieved when μ=0.6. It indicates that the direct lexical information contributes slightly more than the topical information. The results also show the complementarity between these two kinds of information on negation focus identification.

For λ, it has very little, if any, effect on performance, when λ is set from 0.5 to 0.85. It indicates that the contextual information (the first term in Eq.(9)) contributes more than the document information (the second term in Eq.(9)) on negation focus identification.

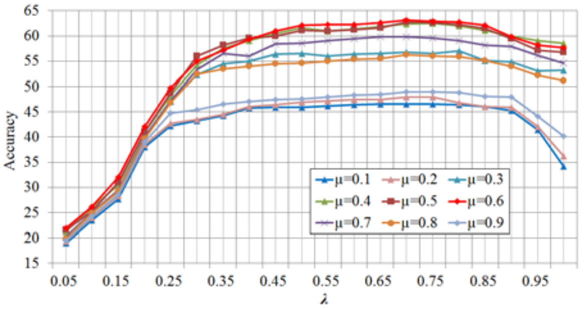

Figure 2. Influence of the damping factors μ and λ.

Moreover, the results also show that these two parameters have little impact in a certain range on performance (μ:0.4~0.6; λ:0.5~0.85), which suggests that the approach is robust to a certain extent. Therefore, we set μ=0.6 and λ=0.7 in the following experiments.

Besides, we also evaluate the other minor parameters in our model. Due to space limit, we do not report all of results here and set parameters to the following values: setting window size s=1 (the previous and next sentences) and the number of topic T=40, adopting the word co-occurrence similarity to calculate the similarity between words, and using dot-product to measure both $P_t(j|i)$ and $R_d(w_i)$.

4.2 Comparison with Other Methods

In the word-topic graph models, two primary improvements are proposed: (a) updating the word transition probability $P_w(j|i)$ by adding a topical layer ("TL"), and (b) assigning a document-sensitive value to word node ("DS").

model	Acc.
WLM	52.61
WTGM (TL)	65.74
WTGM(TL+DS)	69.39

Table 1. Performance of the word-topic graph model.

Table 1 shows that the word-topic graph model (WTGM) is significantly better (+16.78%, p<0.01) than the graph model with only word layer (WLM), which justifies the effectiveness of the topical layer. In addition, the results also indicate that the word-topic graph model not only takes the topical information into account ("TL"), but also considers the semantic relationship in current document ("DS").

We select two supervised baseline methods to compare with our word-topic graph model. One is a decision tree-based system described in Blanco and Moldovan (2011), and the other one is a SVM-based system which takes advantage of both syntactic features and contextual features (Zou et al., 2014).

system	Acc.
B&M(2011)	63.20
Zou et al.(2014)	67.14
Ours	69.39

Table 2. Comparison Results.

Table 2 shows that our word-topic graph model performs significantly better than the two others by 6.19% (p<0.01) and 2.25% (p<0.01), respectively. The results support our viewpoint that the topical information in context can help to find the negation focus, and the word-topic graph model we proposed is effective. Moreover, it is also worth noting that our method is

[3] http://www.cis.upenn.edu/~treebank/
[4] http://gibbslda.sourceforge.net/

unsupervised, which does not need the prior knowledge for training, while the other two supervised baselines employ the golden features, such as the POS tag, constituent tree, and dependency tree.

5 Conclusion

In this paper, we propose an unsupervised word-topic graph model, which represents and measures the word importance by using contextual information from both lexical and topical perspectives. And then, we propose a document-sensitive biased PageRank algorithm to calculate the ranking scores of negation focus candidates. Experimental results show that our method achieves better performance than other baselines without any annotated data.

The main shortcoming of our method is that not all of negation focus can be identified by the context. As our statistics, at least 17% of samples are hard to be determined by human beings when ignoring the information in current sentence. Therefore, in future work, we will focus on investigating an effective method to integrate the local lexical/syntactic information and the global contextual discourse information.

Acknowledgments

This research is supported by the National Natural Science Foundation of China, No.61272260, No.61331011, No.61273320, and the Major Project of College Natural Science Foundation of Jiangsu Province, No.11KJA520003. The authors would like to thank the anonymous reviewers for their insightful comments and suggestions.

Reference

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Eduardo Blanco and Dan Moldovan. 2011. Semantic Representation of Negation Using Focus Detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 581-589, Portland, Oregon, June 19-24, 2011.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022, January.

Tom Griffiths. 2002. Gibbs sampling in the generative model of Latent Dirichlet Allocation. Tech. rep., Stanford University.

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

Jianhua Lin. 1991. Divergence measures based on Shannon entropy. *IEEE Transactions on Information Theory*, 37(14), 145-151.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic Keyphrase Extraction via Topic Decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366-376, MIT, Massachusetts, USA, 9-11 October 2010.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404-411.

George A. Miller. 1995. Wordnet: a lexical database for english. Commun. ACM, 38(11):39-41.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 265-274, Montreal, Canada, June 7-8, 2012.

Lawrence Page, Sergey Brin, Rajeev Motwani, et al. 1998. The pagerank citation ranking: Bringing order to the web. *Technical report*, Stanford University.

Sabine Rosenberg and Sabine Bergler. 2012. UConcordia: CLaC Negation Focus Detection at *Sem 2012. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 294-300, Montreal, Canada, June 7-8, 2012.

Bowei Zou, Qiaoming Zhu, and Guodong Zhou. 2014. Negation Focus Identification with Contextual Discourse Information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 522-530, Baltimore, Maryland, USA, June 23-25 2014.