Topic Models: Accounting Component Structure of Bigrams

Michael Nokel
Lomonosov Moscow State University,
Russian Federation
mnokel@gmail.com

Natalia Loukachevitch
Lomonosov Moscow State University,
Russian Federation
louk_nat@mail.ru

Abstract

The paper describes the results of an empirical study of integrating bigram collocations and similarities between them and unigrams into topic models. First of all, we propose a novel algorithm PLSA-SIM that is a modification of the original algorithm PLSA. It incorporates bigrams and maintains relationships between unigrams and bigrams based on their component structure. Then we analyze a variety of word association measures in order to integrate top-ranked bigrams into topic models. All experiments were conducted on four text collections of different domains and languages. The experiments distinguish a subgroup of tested measures that produce top-ranked bigrams, which demonstrate significant improvement of topic models quality for all collections, when integrated into PLSA-SIM algorithm.

1 Introduction

Topic modeling is one of the latest applications of machine learning techniques to the natural language processing. Topic models identify which topics relate to each document and which words form each topic. Each topic is defined as a multinomial distribution over terms and each document is defined as multinomial distribution over topics (Blei et al., 2003). Topic models have achieved noticeable success in various areas such as information retrieval (Wei and Croft, 2006), including such applications as multi-document summarization (Wang et al., 2009), text clustering and categorization (Zhou et al., 2009), and other natural language processing tasks such as word sense disambiguation (Boyd-Graber et al., 2007), machine translation (Eidelman et al., 2012). Among most well-known models are Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is based on Dirichlet prior distribution, and Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), which is not connected with any parametric prior distribution.

One of the main drawback of the topic models is that they utilize "bag-of-words" model that discards word order and is based on the word independence assumption. There are numerous studies, where the integration of collocations, n-grams, idioms and multi-word terms into topic models is investigated. However, it often leads to a decrease in the model quality due to increasing size of a vocabulary or to a serious complication of the model (Wallach, 2006; Griffiths et al., 2007; Wang et al., 2007).

The paper proposes a novel approach taking into account bigram collocations and relationship between them and unigrams in topic models (such as *citizen – citizen of country – citizen of union – European citizen – state citizen*; *categorization – document categorization – term categorization – text categorization*). This allows us to create a novel method of integrating bigram collocations into topic models that does not consider bigrams being as "black boxes", but maintains the relationship between unigrams and bigrams based on their component structure. The proposed algorithm leads to significant improvement of topic models quality measured in perplexity and topic coherence (Newman et al., 2010) without complications of the model.

All experiments were carried out using PLSA algorithm and its modifications on four corpora of different domains and languages: the English part of Europarl parallel corpus, the English part of JRC-Acquis parallel corpus, ACL Anthology Reference corpus, and Russian banking magazines.

The rest of the paper is organized as follows. In the section 2 we focus on related work. Section 3

proposes a novel algorithm PLSA-SIM that incorporates bigrams and similarities between them and unigrams into topic models. Section 4 describes the datasets used in experiments, all preprocessing steps and metrics used to evaluate the quality. In the section 5 we perform an extensive analysis of a variety of measures for integrating top-ranked bigrams into topic models. And in the last section we draw conclusions.

2 Related Work

The idea of using collocations in topic models is not a novel one. Nowadays there are two kinds of methods proposed to deal with this problem: creation of a unified probabilistic model and preliminary extraction of collocations and n-grams with further integration into topic models.

Most studies belong to the first kind of methods. So, the first movement beyond "bag-of-words" assumption has been made by Wallach (2006), where the Bigram Topic Model was presented. In this model word probabilities are conditioned on the immediately preceding word. The LDA Collocation Model (Griffiths et al., 2007) extends the Bigram Topic Model by introducing a new set of variables and thereby giving a flexibility to generate both unigrams and bigrams. Wang et al. (2007) proposed the Topical N-Gram Model that adds a layer of complexity to allow the formation of bigrams to be determined by the context. Hu et al. (2008) proposed the Topical Word-Character Model challenging the assumption that the topic of an n-gram is determined by the topics of composite words within the collocation. This model is mainly suitable for Chinese language. Johnson (2010) established connection between LDA and Probabilistic Context-Free Grammars and proposed two probabilistic models combining insights from LDA and Adaptor Grammars to integrate collocations and proper names into the topic model.

While all these models have a theoretically elegant background, they are very complex and hard to compute on real datasets. For example, Bigram Topic Model has W^2T parameters, compared to WT for LDA and $WT + DT$ for PLSA, where W is the size of vocabulary, D is the number of documents, and T is the number of topics. Therefore such models are mostly of theoretical interest.

The algorithm proposed in Lau et al. (2013) belongs to the second type of methods that use collocations in topic models. The authors extract bigram collocations via t-test and replace separate units by top-ranked bigrams at the preprocessing step. They use two metrics of topic quality: perplexity and topic coherence (Newman et al., 2010) and conclude that incorporating bigram collocations into topics results in worsening perplexity and improving topic coherence.

Our current work also belongs to the second type of methods and distinguishes from previous papers such as Lau et al. (2013) in that our approach does not consider bigrams as "black boxes", but maintains information about the inner structure of bigrams and relationships between bigrams and component unigrams, which leads to improvement in both metrics: perplexity and topic coherence.

The idea to utilize prior natural language knowledge in topic models is not a novel one. So, Andrzejewski et al. (2009) incorporated domain-specific knowledge by Must-Link and Cannot-Link primitives represented by a novel Dirichlet Forest prior. These primitives control that two words tend to be generated by the same or separate topics. However, this method can result in an exponential growth in the encoding of Cannot-Link primitives and thus has difficulty in processing a large number of constraints (Liu, 2012). Another method of incorporating such knowledge is presented in Zhai (2010) where a semi-supervised EM-algorithm was proposed to group expressions into some user-specified categories. To provide a better initialization for EM-algorithm the method employs prior knowledge that expressions sharing words and synonyms are likely to belong to the same group. Our current work distinguishes from these ones in that we incorporate similarity links between unigrams and bigrams into the topic model in a very natural way counting their co-occurrences in documents. The proposed approach does not increase the complexity of the original PLSA algorithm.

3 PLSA-SIM algorithm

As mentioned above, original topic models utilize the "bag-of-words" assumption that assumes word independence. And bigrams are usually added to topic models as "black boxes" without any ties with other words. So, bigrams are added to the vocabulary as single tokens and in each document containing any of added bigrams the frequencies

of unigram components are decreased by the frequencies of bigrams (Lau et al., 2013). Thus "bag-of-words" assumption holds.

However, there are many similar unigrams and bigrams that share the same lemmas (i.e, *correction – correction of word – error correction – spelling correction*; *rail – rail infrastructure – rail transport – use of rail*) and others in documents. We should note such bigrams do not only have identical words, but many of them maintain semantic and thematic similarity. At the same time other bigrams with the same words (i.e., idioms) can have significant semantic differences. To take into account these different situations, we hypothesized that similar bigrams sharing the same unigram components should often belong to the same topics, if they often co-occur within the same texts.

To verify this hypothesis we precompute sets of similar unigrams and bigrams sharing the same lemmas and propose novel PLSA-SIM algorithm that is the modification of the original PLSA algorithm. We will rely on the description found in Vorontsov and Potapenko (2014) and use the following notations (further in the paper we will use notation "term" when speaking about both unigrams and bigrams):

- D – the collection of documents;
- T – the set of inferred topics;
- W – the vocabulary (the set of unique terms found in the collection D);
- $\Phi = \{\phi_{wt} = p(w|t)\}$ – the distribution of terms w over topics t;
- $\Theta = \{\theta_{td} = p(t|d)\}$ – the distribution of topics t over documents d;
- $S = \{S_w\}$ – the sets of similar terms (S_w is the set of terms similar to w, that is $S_w = \{w \bigcup_v wv \bigcup_v vw\}$, where w is the lemmatized unigram, while wv and vw are lemmatized bigrams);
- n_{dw}, n_{ds} – the number of occurrences of the terms w, s in the document d;
- $\hat{n}_{wt}$ – the estimate of frequency of the term w in the topic t;
- $\hat{n}_{td}$ – the estimate of frequency of the topic t in the document d;
- $\hat{n}_t$ – the estimate of frequency of the topic t in the text collection D;
- n_d – the number of words in the document d.

The pseudocode of PLSA-SIM algorithm is presented in the Algorithm 1. The only modifications of the original algorithm concern lines 6 and 9, where we introduce auxiliary variable f_{dw}, which takes into account pre-computed sets of similar terms. Thus, the weight of such terms is increased within each document.

Algorithm 1: PLSA-SIM algorithm: PLSA with similar terms

Input: collection of documents D, number of topics $|T|$, initial distributions Θ and Φ, sets of similar terms S
Output: distributions Θ and Φ

1 **while** *not meet the stop criterion* **do**
2 **for** $d \in D$, $w \in W$, $t \in T$ **do**
3 $\hat{n}_{wt} = 0$, $\hat{n}_{td} = 0$, $\hat{n}_t = 0$, $n_d = |d|$
4 **for** $d \in D$, $w \in W$ **do**
5 $Z = \sum_t \phi_{wt}\theta_{td}$,
   $f_{dw} = n_{dw} + \sum_{s \in S_w} n_{ds}$
6
7 **for** $t \in T$ **do**
8 **if** $\phi_{wt}\theta_{td} > 0$ **then**
9 $\delta = f_{dw}\phi_{wt}\theta_{td}/Z$
10 $\hat{n}_{wt} = \hat{n}_{wt} + \delta$, $\hat{n}_{td} = \hat{n}_d + \delta$,
   $\hat{n}_t = \hat{n}_t + \delta$
11 **for** $w \in W$, $t \in T$ **do**
12 $\phi_{wt} = \hat{n}_{wt}/\hat{n}_t$
13 **for** $d \in D$, $t \in T$ **do**
14 $\theta_{td} = \hat{n}_{td}/n_d$

So, if similar unigrams and bigrams co-occur within the same document, we try to carry them to the same topics. We consider such terms having semantic and thematic similarities. However, if unigrams and bigrams from the same set S_w do not co-occur within the same document, we do no modifications to the original algorithm PLSA. We consider such terms having semantic differences.

4 Datasets and Evaluation

4.1 Datasets and Preprocessing

In our experiments we used English and Russian text collections obtained from different sources:

- For the English part of our study we took three different collections:
  - Europarl multilingual parallel corpus. It was extracted from the proceedings of the European Parliament (`http:`

//www.statmt.org/europarl). The English part includes almost 54 mln. words and 9672 documents.

  - JRC-Acquis multilingual parallel corpus. It represents selected texts of the EU legislation written between the 1950s and 2005 (http://ipsc.jrc.ec.europa.eu/index.php?id=198). The English part contains almost 45 mln. words and 23545 documents.
  - ACL Anthology Reference Corpus. It contains scholarly publications about Computational Linguistics (http://acl-arc.comp.nus.edu.sg/). The corpus includes almost 42 mln. words and 10921 documents.

- For the Russian part of our study we took 10422 Russian articles from several economics-oriented magazines such as Auditor, RBC, Banking Magazine, etc. These documents contain almost 18.5 mln. words.

At the preprocessing step documents were processed by morphological analyzers. For the English corpus we used Stanford CoreNLP tools (http://nlp.stanford.edu/software/corenlp.shtml), while for the Russian corpus we used our own morphological analyzer. All words were lemmatized. We consider only Adjectives, Nouns, Verbs and Adverbs since function words do not play significant role in forming topics. Besides, we excluded words occurring less than five times per the whole text collection.

In addition, we extracted all bigrams in forms of *Noun + Noun*, *Adjective + Noun* and *Noun + of + Noun* for all English collections, and *Noun + Noun in Genitive* and *Adjective + Noun* for the Russian collection. We consider only such bigrams since topics are mainly identified by nouns and noun groups (Wang et al., 2007).

4.2 Evaluation Framework

As for the inferred topics quality, we consider four different intrinsic measures. The first measure is **Perplexity** since it is the standard criterion of topic models quality (Daud et al., 2010):

$$Perplexity(D) = \exp\left\{-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right\}, \quad (1)$$

where n is the number of all considered words in the collection, D is the set of documents in the collection, n_{dw} is the number of occurrences of the word w in the document d, $p(w|d)$ is the probability of appearing the word w in the document d.

The less the value of perplexity is the better the model predicts words w in documents D. Although there were numerous studies arguing that perplexity is not suited to topic model evaluation (Chang et al., 2009; Newman et al., 2010), it is still commonly used for comparing different topic models. Since it is well-known that perplexity computed on the same training collection is susceptible to over-fitting and can give optimistically low values (Blei et al., 2003) we use the standard method of computing hold-out perplexity described in Asuncion et al. (2009). In our experiments we split the collections randomly into the training sets D, on which models are trained, and the validation sets D', on which hold-out perplexity is computed.

Another method of evaluating topic model quality is using **expert opinions**. We provided annotators with inferred topics from the same text collections and instructed them to decide whether the topic was to some extent coherent, meaningful and interpretable. The indicator of topic usefulness is the ease by which one could think of a short label to describe a topic (Newman et al., 2010). In the Table 1 we present incoherent topic that can not be given any label and coherent one with label given by experts.

Top words from topic	Label
have, also, commission, state, more, however	–
vessel, fishing, fishery, community, catch, board	*fishing*

Table 1: Examples of incoherent and coherent topics

Since involving experts is time-consuming and expensive, there were several attempts to propose a method for automatic evaluation of topic models quality that would go beyond perplexity and would be correlated with expert opinions. The formulation of such a problem is very complicated since experts can quite strongly disagree with each other. However, it was recently shown that it is possible to evaluate topic coherence automatically using word semantics with precision, almost coinciding with experts (Newman et al., 2010; Mimno et al., 2011). The proposed metric measures interpretability of topics based on human judge-

ment (Newman et al., 2010). As topics are usually presented to users via their top-N topic terms, the *topic coherence* evaluates whether these top terms correspond to the topic or not. Newman et al. (2010) proposed an automated variation of the coherence score based on pointwise mutual information (**TC-PMI**):

$$TC\text{-}PMI(t) = \sum_{j=2}^{10} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_j)P(w_i)}, \quad (2)$$

where $(w_1, w_2, \ldots, w_{10})$ are the top-10 terms in a topic, $P(w_i)$ and $P(w_j)$ are probabilities of unigrams w_i and w_j respectively, while $P(w_j, w_i)$ is the probability of bigram (w_j, w_i). The final measure of topic coherence is calculated by averaging $TC\text{-}PMI(t)$ measure by all topics t.

This score is proven to demonstrate high correlation with human judgement (Newman et al., 2010). The proposed metric considers only top-10 words in each topic since they usually provide enough information to form the subject of the topic and distinguishing features from other topics. Topic coherence is becoming more widely used to evaluate topic model quality along with perplexity. For example, Stevens et al. (2012) showed that this metric is strongly correlated with expert estimates. Also Andrzejewski et al. (2011) simply used it for evaluating topic model quality.

Following the approach proposed by Mimno et al. (2011) we compute probabilities by dividing the number of documents where the unigram or bigram occurred by the number of documents in the collection. To avoid optimistically high values we use external corpus for this purpose – namely, Russian and English Wikipedia. We should note that we do not consider another variation of topic coherence based on log conditional probability (*TC-LCP*) proposed by Mimno et al. (2011) since it was shown in Lau et al. (2013) that it works significantly worse than *TC-PMI*.

We should note that while incorporating the knowledge of similar unigrams and bigrams into topic models in the proposed algorithm, we encourage such terms to be in the top-10 terms in inferred topics. Therefore, we increase TC-PMI metric unintentionally since such terms are likely to co-occur within the same documents. So we decided to use also modification of this metric to consider not top-10 terms in topics but top-10 non-similar terms in topics (this metric will be further called as **TC-PMI-nSIM**).

5 Integrating bigrams into topic models

To compare proposed algorithm with the original one we extracted all bigrams found in each document of collections. For ranking bigrams we utilized *Term Frequency (TF)* or one of the following 19 word association measures:

1. *Mutual Information (MI)* (Church and Hanks, 1990);
2. *Augmented MI* (Zhang, 2008);
3. *Normalized MI* (Bouma, 2009);
4. *True MI* (Deane, 2005);
5. *Cubic MI* (Daille, 1995);
6. *Symmetric Conditional Probability* (Lopes and Silva, 1999);
7. *Dice Coefficient (DC)* (Smadja et al., 1996);
8. *Modified DC* (Kitamura and Matsumoto, 1996);
9. *Lexical Cohesion* (Park et al., 2002);
10. *Gravity Count* (Daudarvičius and Marcinkevičienė, 2003);
11. *Simple Matching Coefficient* (Daille, 1995);
12. *Kulczinsky Coefficient* (Daille, 1995);
13. *Ochiai Coefficient* (Daille, 1995);
14. *Yule Coefficient* (Daille, 1995);
15. *Jaccard Coefficient* (Jaccard, 1901);
16. *T-Score*;
17. *Z-Score*;
18. *Chi Square*;
19. *Loglikelihood Ratio* (Dunning, 1993).

According to the results of Lau et al. (2013) we decided to integrate top-1000 bigrams into all topic models under consideration. We should note that in all experiments described in the paper we fixed the number of topics and the number of iterations of algorithms to 100.

We conducted experiments with all **20** aforementioned measures on all four text collections in order to compare the quality of the original algorithm PLSA, PLSA with top-1000 bigrams added as "black boxes", and PLSA-SIM algorithm with the same top-1000 bigrams.

According to the results of experiments we have revealed two groups of measures.

The first group contains *MI*, *Augmented MI*, *Normalized MI*, *DC*, *Chi-Square*, *Symmetrical Conditional Probability*, *Simple Matching Coefficient*, *Kulczinsky Coefficient*, *Yule Coefficient*, *Ochiai Coefficient*, *Jaccard Coefficient*, *Z-Score*, and *Loglikelihood Ratio*. We got nearly the same levels of perplexity and topic coherence when top

bigrams ranked by these measures were integrated into all tested topic models. This is explained by the fact that these measures rank up very special, non-typical and low frequency bigrams. In the Table 2 we present results of integrating top-1000 bigrams ranked by *MI* for all text collections.

Corpus	Model	Perplexity	TC-PMI	TC-PMI-nSIM
Banking	*PLSA*	1724.2	86.1	86.1
	PLSA + bigrams	1714.1	84.2	84.2
	PLSA-SIM + bigrams	1715.4	84.1	84.1
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA + bigrams	1584.6	55	55
	PLSA-SIM + bigrams	1591.3	55.2	55.2
JRC	*PLSA*	812.1	67	67
	PLSA + bigrams	815.4	66.3	66.3
	PLSA-SIM + bigrams	815.6	66.4	66.4
ACL	*PLSA*	2134.7	74.8	74.8
	PLSA + bigrams	2138.1	75.5	75.5
	PLSA-SIM + bigrams	2144.8	75.8	75.8

Table 2: Results of integrating top-1000 bigrams ranked by MI into topic models

The second group includes *TF*, *Cubic MI*, *True MI*, *Modified DC*, *T-Score*, *Lexical Cohesion* and *Gravity Count*. We got worsened perplexity and improved topic coherence, when top bigrams ranked by these measures were integrated into PLSA algorithm as "black boxes". But when they were used in PLSA-SIM topic models, it led to significant improvement of all metrics under consideration. This is explained by the fact that these measures rank up high frequent, typical bigrams. In the Table 3 we present results of integrating top-1000 bigrams ranked by *TF* for all text collections.

So, we succeed to achieve better quality for both languages using the proposed algorithm and the second group of measures.

For the expert evaluation of topic model quality we invited two linguistic experts and gave them topics inferred by the original PLSA algorithm and by the proposed PLSA-SIM algorithm with top-1000 bigrams ranked by TF (term frequency). The task was to classify given topics into 2 classes: whether they can be given a subject name (we will further mark such topics as '+') or not (we will further mark such topics as '–'). In the Table 4 we present results for all text collections except ACL Anthology Reference Corpus because for the correct markup advance knowledge in computational linguistics is required.

Corpus	Model	Perplexity	TC-PMI	TC-PMI-nSIM
Banking	*PLSA*	1724.2	86.1	86.1
	PLSA + bigrams	2251.8	98.8	98.8
	PLSA-SIM + bigrams	**1450.6**	**156.5**	**102.6**
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA + bigrams	1993.5	57.3	57.3
	PLSA-SIM + bigrams	**1431.6**	**127.7**	**84.7**
JRC	*PLSA*	812.1	67	67
	PLSA + bigrams	1038.9	72	72
	PLSA-SIM + bigrams	**743.7**	**108.4**	**76.9**
ACL	*PLSA*	2134.7	74.8	74.8
	PLSA + bigrams	2619.3	73.7	73.7
	PLSA-SIM + bigrams	**1806.4**	**152.7**	**87.8**

Table 3: Results of integrating top-1000 bigrams ranked by TF into topic models

Corpus	Model	Expert 1		Expert 2	
		+	–	+	–
Banking	*PLSA*	93	7	92	8
	PLSA + bigrams	92	8	95	5
	PLSA-SIM + bigrams	**95**	**5**	**97**	**3**
JRC	*PLSA*	92	8	90	10
	PLSA + bigrams	94	6	97	3
	PLSA-SIM + bigrams	**97**	**3**	**100**	**0**
Europarl	*PLSA*	97	3	99	1
	PLSA + bigrams	95	5	99	1
	PLSA-SIM + bigrams	**98**	**2**	**100**	**0**

Table 4: Results of expert markup of topics

As we can see, in the case of PLSA-SIM algorithm with top-1000 bigrams ranked by TF the amount of inferred topics, for which labels can be given, is increased for all text collections. It is also worth noting that adding bigrams as "black boxes" does not increase the amount of such inferred topics. This result also confirms that the proposed algorithm improves the quality of topic models.

In the Table 5 we present top-5 words from one random topic for each corpus for original

PLSA and PLSA-SIM algorithms with top-1000 bigrams ranked by TF. Within each text collection we present topics discussing the same subject.

Banking		**Europarl**	
PLSA	PLSA-SIM	PLSA	PLSA-SIM
Banking	*Financial system*	*Financial*	*Economic crisis*
Bank	*Financial market*	*Crisis*	*Financial crisis*
Sector	*Financial sector*	*Have*	*European economy*
Financial	*Financial*	*European*	*Time of crisis*
System	*Financial institute*	*Market*	*Crisis*
JRC-Acquis		**ACL**	
PLSA	PLSA-SIM	PLSA	PLSA-SIM
Transport	*Transport*	*Tag*	*Tag*
Road	*Transport service*	*Word*	*Tag set*
Nuclear	*Road transport*	*Corpus*	*Tag sequence*
Vehicle	*Transport sector*	*Tagger*	*Unknown word*
Material	*Air transport*	*Tagging*	*Speech tag*

Table 5: Top-5 words from topics inferred by PLSA and PLSA-SIM algorithms

We should note that we used only intrinsic measures of topic model quality in the paper. In the future we would like to test improved topic models in such applications of information retrieval as text clustering and categorization.

6 Conclusion

The paper presents experiments on integrating bigrams and similarities between them and unigrams into topic models. At first, we propose the novel algorithm PLSA-SIM that incorporates similar unigrams and bigrams into topic models and maintains relationships between bigrams and unigram components. The experiments were conducted on the English parts of Europarl and JRC-Acquis parallel corpora, ACL Anthology Reference corpus and Russian banking articles distinguished two groups of measures ranking bigrams. The first group produces top bigrams, which, if added to topic models either as "black boxes" or not, results in nearly the same quality of inferred topics. However, the second group produces top bigrams, which, if added to the proposed PLSA-SIM algorithm, results in significant improvement in all metrics under consideration.

Acknowledgements

This work is partially supported by RFBR grant N14-07-00383.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. *Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors*. Proceedings of the 26th Annual International Conference on Machine Learning: 25–32.

David Andrzejewski and David Buttler. 2011. *Latent Topic Feedback for Information Retrieval*. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 600–608.

Arthur Asuncion, Max Welling, Padhraic Smyth, Yee Whye Teh. 2009. *On Smoothing and Inference for Topic Models*. Proceedings of the 25th International Conference on Uncertainty in Artificial Intelligence: 27–34.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. *Latent Dirichlet Allocation*. Journal of Machine Learning Research, volume 3: 993–1022.

Gerlof Bouma. 2009. *Normalized (Pointwise) Mutual Information*. Proceedings of the Biennial GSCL Conference: 31–40.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. *A Topic Model for Word Sense Disambiguation*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning: 1024–1033.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David M. Blei. 2009. *Reading Tea Leaves: How Human Interpret Topic Models*. Proceedings of the 24th Annual Conference on Neural Information Processing Systems: 288–296.

Kenneth Ward Church, and Patrick Hanks. 1990. *Word Association Norms, Mutual Information, and Lexicography*. Computational Linguistics, volume 16: 22–29.

Beatrice Daille. 1995. *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering* PhD Dissertation. University of Paris, Paris.

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad. 2010. *Knowledge discovery through directed probabilistic topic models: a survey*. Frontiers of Computer Science in China, 4(2): 280–301.

Vidas Daudarvičius and Rūta Marcinkevičienė. 2003. *Gravity Counts for the Boundaries of Collocations*. International Journal of Corpus Linguistics, 9(2): 321–348.

Paul Deane. 2005. *A Nonparametric Method for Extraction of Candidate Phrasal Terms*. Proceedings of the 43rd Annual Meeting of the ACL: 605–613.

Ted Dunning. 1993. *Accurate Methods for the Statistics of Surprise and Coincidence*. International Journal of Computational Linguistics, 19(1): 61–74.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. *Topic Models for Dynamic Translation Model Adaptation*. Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics, volume 2: 115–119.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. *Topics in Semantic Representation*. Psychological Review, 114(2): 211–244.

Thomas Hofmann. 1999. *Probabilistic Latent Semantic Indexing*. Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval: 50–57.

Wei Hu, Nobuyuki Shimizu, Hiroshi Nakagawa, and Huanye Shenq. 2008. *Modeling Chinese Documents with Topical Word-Character Models*. Proceedings of the 22nd International Conference on Computational Linguistics: 345–352.

Paul Jaccard. 1901. *Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines*. Bull. Soc. Vaudoise sci. Natur. V. 37. Bd. 140: 241–272.

Mark Johnson M. 2010. *PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names*. Proceedings of the 48th Annual Meeting of the ACL: 1148–1157.

Mihoko Kitamura, and Yuji Matsumoto. 1996. *Automatic Extraction of Word Sequence Correspondences in Parallel Corpora*. Proceedings of the 4th Annual Workshop on Very Large Corpora: 79–87.

Jey Han Lau, Timothy Baldwin, and David Newman. 2013. *On Collocations and Topic Models*. ACM Transactions on Speech and Language Processing, 10(3): 1–14.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Jose Gabriel Pereira Lopes, and Joaquim Ferreira da Silva. 1999. *A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units*. Proceedings of the 6th Meeting on the Mathematics of Language: 369–381.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, Andrew McCallum. 2011. *Optimizing Semantic Coherence in Topic Models*. Proceedings of EMNLP'11: 262–272.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. *Automatic Evaluation of Topic Coherence*. Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: 100–108.

Youngja Park, Roy J. Byrd, and Branimir K. Boguraev. 2002. *Automatic Glossary Extraction: Beyond Terminology Identification*. Proceedings of the 19th International Conference on Computational Linguistics: 1–7.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. Computational Linguistics, 22(1): 1–38.

Keith Stevens, Philip Kegelmeyer, David Adnrzejewski, and David Buttler. 2012. *Exploring Topic Coherence over Many Models and Many Topics*. Proceedings of EMNLP-CoNLL'12: 952–961.

Konstantin V. Vorontsov, and Anna A. Potapenko. 2014. *Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization*. Proceedings of AIST'2014. LNCS, Springer Verlag-Germany, volume CCIS 439: 28–45.

Hanna M. Wallach. 2006. *Topic Modeling: Beyond Bag-of-Words*. Proceedings of the 23rd International Conference on Machine Learning: 977–984.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. *Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval*. Proceedings of the 2007 Seventh IEEE International Conference on Data Mining: 697–702.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-Document Summarization using Sentence-based Topic Models. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers: 297–300.

Xing Wei and W. Bruce Croft. 2006. *LDA-Based Document Models for Ad-hoc Retrieval*. Proceedings of the 29th International Conference on Research and Development in Information Retrieval: 178–185.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2010. *Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints*. Proceedings of the 23rd International Conference on Computational Linguistics: 1272–1280.

Wen Zhang, Taketoshi Yoshida, Tu Bao Ho, and Xijin Tang. 2008. *Augmented Mutual Information for Multi-Word Term Extraction*. International Journal of Innovative Computing, Information and Control, 8(2): 543–554.

Shibin Zhou, Kan Li, and Yushu Liu. 2009. *Text Categorization Based on Topic Model*. International Journal of Computational Intelligence Systems, volume 2, No. 4: 398–409.