A Novel Topic Model for Automatic Term Extraction

Sujian Li[1] Jiwei Li[1] Tao Song[1] Wenjie Li[2] Baobao Chang[1]
[1] Key Laboratory of Computational Linguistics (Peking University), Ministry of Education,
School of Electronics Engineering and Computer Science, CHINA
[2] The Innovative Intelligent Computing Center,
The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, CHINA
{lisujian, bdlijiwei, stao, chbb} @pku.edu.cn; cswjli@comp.polyu.edu.hk

ABSTRACT
Automatic term extraction (ATE) aims at extracting domain-specific terms from a corpus of a certain domain. *Termhood* is one essential measure for judging whether a phrase is a term. Previous researches on *termhood* mainly depend on the word frequency information. In this paper, we propose to compute *termhood* based on semantic representation of words. A novel topic model, namely i-SWB, is developed to map the domain corpus into a latent semantic space, which is composed of some general topics, a background topic and a documents-specific topic. Experiments on four domains demonstrate that our approach outperforms the state-of-the-art ATE approaches.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing, Thesauruses.*

General Terms
Algorithms, Experimentation.

Keywords
Term Extraction, Topic Model, Termhood.

1. INTRODUCTION
So far, most researches on automatic term extraction have been guided by two essential measures defined by [6], namely *unithood* and *termhood*. *Unithood* examines syntactic formation of terms or the degree (or significance) of the association among the term constituents. *Termhood*, on the other hand, aims to capture the semantic relatedness of a term to a domain concept. However, there is no uniform definition of what is semantic relatedness, and how to compute *termhood* is still an open problem.

Previous researches have attempted to measure *termhood* by applying several statistical measures within a domain or across domains, such as TF-IDF, C-value/NC-value [5], co-occurrence [4] and inter-domain entropy [2]. These statistical measures often ignore the informative words with very high frequency or very low frequency and do not take into account the semantics carried by terms. Taking the term "*NRZ electrical input*" in the *electric engineering* domain for example, "*NRZ*" only occurs in a few documents while "*electrical*" occurs in many documents frequently. Using TF-IDF to measure *termhood*, low scores will be assigned to both "*NRZ*" and "*electrical*", which in turn causes the term "*NRZ electrical input*" to have a low *termhood*. It is obvious that frequency-based measures will keep many real terms out of the door. In fact, a domain is described semantically from various aspects. Again, let's take the *electric engineering* domain for example. The words like "*input*" emphasize some specific topic in a domain while the words like "*electrical*" provide the background of that domain. There also exist a cluster of words like "*NRZ*" which occur in the corpus infrequently, but tend to occur in a few documents frequently. Such words can reflect some special characteristics of the domain. Based on these observations, we argue that three semantic aspects can be used in the representation of words: *Domain background* words (e.g. *electrical*) describe the domain in general. *Domain topic* words (e.g. *input*) represent a certain topic in a given domain. *Domain documents-specific* words (e.g. *NRZ*) are specific to a small number of documents and exhibit the characteristics of the domain. We assume that a term can be recognized by identifying whether its constituent words belong to some of the three semantic aspects.

As for semantic representation of words, unsupervised topic models have shown their advantages [1] [3]. Latent Dirichlet Allocation (LDA) is a well-known example of such models. It posits that each document can be seen as a mixture of latent topics and each topic as the distribution over a given vocabulary. To trade-off generality and specificity of words, Chemudugunta et al. [3] further defined the special words with background (SWB) model that allowed words to be modeled as originating from general topics, or document-specific topics, or a corpus-wide background topic. The existing work proves that topic models are competent for the semantic representation of words. However, to our knowledge, no prior work has introduced such kind of semantic representation to term extraction.

Inspired by Chemudugunta's idea of generality and specificity [3], in this paper we propose a novel topic model, namely **i-SWB** to model the three suggested semantic aspects. In i-SWB, three kinds of topics, namely **background topic**, **general topics**, and **documents-specific topic** are correspondingly constructed to generate the words in a domain corpus. Compared with Chemudugunta's SWB model, there are two main improvements in i-SWB to tailor to term extraction. First, specificity in i-SWB is modeled at the corpus level and one documents-specific topic is set to identify a cluster of idiosyncratic words from the whole corpus. Thus, i-SWB avoids the computationally intensive problem in SWB where the number of document-specific topics grows linearly with the number of documents. Second, i-SWB makes use of both document frequency (DF) and topic information to control the generation of words, while SWB only uses a simple multinomial variable to control which topic a word is generated from. This improvement comes from the following

Conference'10, Month 1–2, 2010, City, State, Country.

findings that have been verified in the experiments: the words occurring in many documents and distributing over many general topics with higher probability in LDA are likely to present background information, while the words occurring in a few documents only and distributing over a certain topic in LDA with medium or low probabilities are usually idiosyncratic and provide some special information of the domain. Next, with the semantic representation of words in i-SWB, we implement an ATE system which outperforms the existing ATE approaches.

2. MODEL DESCRIPTION

2.1 LDA and SWB Fundamentals

The hierarchical Bayesian LDA models the probability of a corpus on hidden topics as in Fig. 1(a). The topic distribution of each document θ_d is drawn from a prior Dirichlet distribution $Dir(\alpha)$, and each document word $w_{d,n}$ is sampled from a topic-word distribution ϕ^z specified by a drawn from the topic-document distribution θ_d. The topic assignments z for each word in the corpus can be efficiently sampled via Gibbs sampling, and the predictive distributions for θ and ϕ can be computed by averaging over multiple samples.

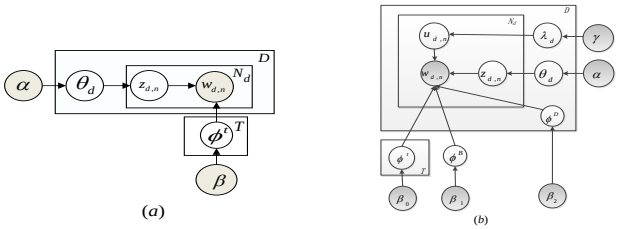

Figure 1. Graphical models for (*a*) LDA and (*b*) SWB.

To formulate background and special information, Chemudugunta et al. (2006) proposed the SWB model as illustrated in Fig. 1(b). In SWB, the variable $u_{d,n}$ acts as a switch: if $u_{d,n}$ = 0, the current word is generated by the general topics as in LDA, whereas if $u_{d,n}$=1 or $u_{d,n}$=2, words are sampled from a corpus specific multinomial or a document-specific multinomial (with symmetric Dirichlet priors β_1 and β_2) respectively. $u_{d,n}$ is sampled from a document-specific multinomial λ_d , and it has a symmetric Dirichlet prior γ. Applying a Gibbs sampler on SWB, we can get the sampling equations for each word $w_{d,n}$ in document d:

$$p(\mu_{d,n}=0, \mathbf{z}_{d,n}=t \mid \mathbf{w}, \mathbf{x}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \alpha, \beta_0, \gamma) \propto \frac{N_{d0,-(d,n)}+\gamma}{N_{d,-(d,n)}+3\gamma} \cdot \frac{C^{TD}_{td,-(d,n)}+\alpha}{\sum_{t'} C^{TD}_{t'd,-(d,n)}+T\alpha} \cdot \frac{C^{WT}_{wt,-(d,n)}+\beta_0}{\sum_{w} C^{WT}_{w't,-(d,n)}+V\beta_0} \quad (1)$$

$$p(\mu_{d,n}=1 \mid \mathbf{w}, \mathbf{x}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \beta_1, \gamma) \propto \frac{N_{d1,-(d,n)}+\gamma}{N_{d,-(d,n)}+3\gamma} \cdot \frac{C^{WB}_{wb,-(d,n)}+\beta_1}{\sum_{w'} C^{WB}_{w'b,-(d,n)}+V\beta_1} \quad (2)$$

$$p(\mu_{d,n}=2 \mid \mathbf{w}, \mathbf{x}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \beta_2, \gamma) \propto \frac{N_{d2,-(d,n)}+\gamma}{N_{d,-(d,n)}+3\gamma} \cdot \frac{C^{WD}_{wd,-(d,n)}+\beta_2}{\sum_{w'} C^{WD}_{w'd,-(d,n)}+V\beta_2} \quad (3)$$

where $-(d,n)$ indicates that the current word $w_{d,n}$ is excluded for the count, V denotes the vocabulary size, N_d is the number of the words in document d and N_{d0}, N_{d1}, and N_{d2} are the number of the words in document d assigned to the latent topics, background topic and document topics, respectively. C^{TD}_{td} is the number of the words that are assigned topic t in document d. C^{WT}_{wt}, C^{WB}_{wb} and C^{WD}_{wd} are the number of the words that $w_{d,n}$ is assigned to topic t, to the background topic and to the documents-specific topic respectively.

2.2 i-SWB Model

In i-SWB, the documents-specific topic is defined at the corpus level and the corpus is composed of three kinds of topics including background topic, documents-specific topic and general topics. To control the generation of these topics, a simple way is to set a variable (e.g. λ_d in SWB) which is drawn from a symmetric Dirichlet prior. As a rule of thumb, document frequency (DF) information can be used as a determining factor to examine the specificity or generality of words related to a domain. In addition, we use the word distribution in general topics as another factor to determine the specificity and generality of words. Fig. 2 and 3 show the graphic model and the generation process of i-SWB. Instead of λ_d, a three-dimensional vector $\pi_{d,n}$ is used to control topic generation and its value is determined by an experience function $p(DF_{w_{d,n}}, \Phi_{w_{d,n}})$ where $DF_{w_{d,n}}$ denotes the document frequency of the word $w_{d,n}$ and $\Phi_{w_{d,n}}$ denotes the probability vector that the word $w_{d,n}$ distributes over all the general topics.

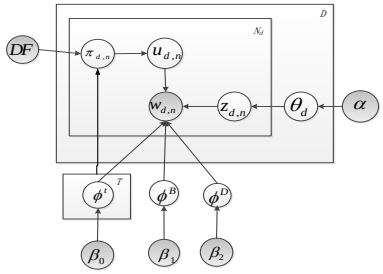

Figure 2. Graphical model of i-SWB.

1. Draw $\phi^B \sim Dir(\beta_1)$, $\phi^D \sim Dir(\beta_2)$
2. For each topic $t \in [1,T]$:
   draw $\phi^t \sim Dir(\beta_0)$
3. For each document $d \in [1,D]$:
   (a) draw $\theta_d \sim Dir(\alpha)$
   (b) For each word $w_{d,n}$:
      i. draw $\pi_{d,n} \sim p(DF_{w_{d,n}}, \Phi_{w_{d,n}})$
      ii. draw $\mu_{d,n} \sim multi(\pi_{d,n})$
      iii. draw:
         if $\mu_{d,n}=2$, draw $w_{d,n} \sim multi(\phi^D)$
         if $\mu_{d,n}=1$, draw $w_{d,n} \sim multi(\phi^B)$
         if $\mu_{d,n}=0$, draw $z_{d,n} \sim multi(\theta_d)$
         and $w_{d,n} \sim multi(\phi^{z_{d,n}})$

Figure 3. Generation process of i-SWB.

To formally present the i-SWB model, let V be the vocabulary size and D be the number of documents. There are T general topics $\phi^t (1 \le t \le T)$, one background topic ϕ^B and one documents-specific topic ϕ^D which have symmetric Dirichlet priors of β_0, β_1 and β_2 respectively. Each topic is characterized by a distribution over the V words. α is the fixed parameter of symmetric Dirichlet prior for the D document-topic multinomials represented by a $D \times T$ matrix θ. Let $w_{d,n}$ be the observed variable representing the n^{th} word in document d, $u_{d,n}$ the hidden variable denoting which kind of topic $w_{d,n}$ is assigned to, and $z_{d,n}$ the hidden variable indicating that the general topic $w_{d,n}$ may be assigned to.

2.3 Model Inference

We use a Gibbs sampler to perform model inference. Due to space limitation, we give the result of Gibbs Sampling directly.

$$P(u_{d,n}=0, z_{d,n}=t \mid \mathbf{w}, \mathbf{z}_{-(d,n)}, \pi_{d,n}, \mathbf{u}_{-(d,n)},) \propto \pi^{(0)}_{d,n} \cdot \frac{C^{TD}_{td,-(d,n)}+\alpha}{\sum_{t'} C^{TD}_{t'd,-(d,n)}+T\alpha} \cdot \frac{C^{WT}_{wt,-(d,n)}+\beta_0}{\sum_{w'} C^{WT}_{w't,-(d,n)}+V\beta_0} \quad (4)$$

$$P(u_{d,n}=1 \mid \mathbf{w}, \mathbf{z}, \pi_{d,n}, u_{-(d,n)},) \propto \pi^{(1)}_{d,n} \cdot \frac{C^{WB}_{wb,-(d,n)}+\beta_1}{\sum_{w'} C^{WB}_{w'b,-(d,n)}+V\beta_1} \quad (5)$$

$$P(u_{d,n}=2 \mid \mathbf{w}, \mathbf{z}, \pi_{d,n}, u_{-(d,n)}) \propto \pi^{(2)}_{d,n} \cdot \frac{C^{WD}_{wd,-(d,n)}+\beta_2}{\sum_{w'} C^{WD}_{w'd,-(d,n)}+V\beta_2} \quad (6)$$

Eqs. (4), (5) and (6) are similar to Eqs. (1), (2) and (3), except the first terms in each of them. Then, we determine $\pi_{d,n}$, whose value depends on $DF_{w_{d,n}}$ and $\Phi_{w_{d,n}}$ as follows:

$$\pi^{(0)}_{d,n} = (1-E(\Phi_{w_{d,n}})) \cdot \frac{DF_{w_{d,n}}}{D}; \quad \pi^{(1)}_{d,n} = E(\Phi_{w_{d,n}});$$

$$\pi^{(2)}_{d,n} = (1-E(\Phi_{w_{d,n}})) \cdot (1-\frac{DF_{w_{d,n}}}{D}) \quad (7)$$

where $E(\Phi_{w_{d,n}})$ is used to measure the evenness of distribution that $w_{d,n}$ is assigned to each general topic.

$$E(\Phi_{w_{d,n}}) = \frac{-\sum_{t\in T} \frac{C^{WT}_{w_{d,n}t}}{\sum_{t'} C^{WT}_{w_{d,n}t'}} \log \frac{C^{WT}_{w_{d,n}t}}{\sum_{t'} C^{WT}_{w_{d,n}t'}}}{\log T} \quad (8)$$

where $C^{WT}_{w_{d,n}t}$ denotes the number of times that $w_{d,n}$ is assigned to the general topic *t*. In Eq. (8), the numerator computes the entropy that the word $w_{d,n}$ is assigned to the general topics, and the denominator denotes the maximum entropy value that $w_{d,n}$ is evenly assigned to each general topic. The range of $E(\Phi_{w_{d,n}})$ is (0,1]. The experience Equations (7) and (8) reflect that a background word is more likely to uniformly distribute on general topics and a documents-specific word is more likely to have a larger *DF* value. With one Gibbs sampling, we can also make the following estimation:

$$\theta^t_d = \frac{C^{TD}_{td}+\alpha}{\sum_{t'} C^{TD}_{t'd}+T\alpha}, \ \phi^t_w = \frac{C^{WT}_{wt}+\beta_0}{\sum_{w'} C^{WT}_{w't}+V\beta_0},$$

$$\phi^B_w = \frac{C^{WB}_{wb}+\beta_1}{\sum_{w'} C^{WB}_{w'b}+V\beta_1}, \ \phi^D_w = \frac{C^{WD}_{wd}+\beta_2}{\sum_{w'} C^{WD}_{w'd}+V\beta_2} \quad (9)$$

In our experiments, we set T=20, α=50/T, β_0=0.1, β_1=0.01 and β_2=0.01. We initialize the corpus by sampling each word from the general topics and run 1000 iterations to stabilize the distribution of $\boldsymbol{z}$ and $\boldsymbol{u}$.

3. TERM EXTRACTION

This section will introduce the proposed i-SWB based term extraction. As stated in Section 1, the ATE process is usually guided by two measures: *unithood* for acquiring term candidates and *termhood* for further identifying terms from the candidates. Following the work of Frantzi [5], the *unithood* technique adopted in our work is mainly based on a linguistic filter as the following three steps:

Step 1: Part-of-speech (POS) tagging: We use a Maximum-entropy POS tagger[1] implemented by Standford NLP group to assign a grammatical tag (e.g. noun, verb, adjective etc.) to each word in the corpus.

[1] http://www-nlp.stanford.edu/software/tagger.shtml

Step 2: Linguistic filtering: Since most terms are noun phrases which consist of nouns, adjectives and some variants of verbs and end with a noun word, here we present our method with the filter: (*FW* | *Verb* | *Adj* | *Noun*)* *Noun*, where *FW* usually denotes the unknown words.

Step 3: Frequency recording: for each term candidate, we record its frequency occurring in the whole corpus and exclude those occurring only once from the candidate list.

Now we get a list of term candidates with their frequency $\{c_i, tf_i\}$. Suppose each candidate c_i consists of L_i words $w_{i1}w_{i2}...w_{iL_i}$. To compute *termhood*, each candidate is scored according to tf_i and the results of i-SWB model. According to the values of ϕ^t, ϕ^B and ϕ^D in Eqs. (9), we extract the top H (e.g. 200) highest distributed words for each topic, namely V_t, V_B and V_D, which can be seen as the typical words of the corresponding topic. An intuitive idea is that, a good candidate should be composed of typical words which are representative of a certain topic. Thus, the *termhood* of c_i is computed as:

$$Termhood(c_i) = \log(tf_i) \cdot \sum_{1\le j\le L_i, w_j \in \cup\{V_t\}_{t\in T\cup\{B,D\}}} \phi^{mt_{w_j}}_{w_j}$$

$$mt_{w_j} = \arg\max_{t\in T\cup\{B,D\}} (\phi^t_{w_j}) \quad (10)$$

where mt_{w_j} denotes the topic which w_j is most likely to be assigned. To compute c_i, we only consider those constituent words that are typical of a certain topic. Then the candidates with the highest *termhood* values are taken as terms.

4. EXPERIMENTS

4.1 Experiment Setup

We evaluate the i-SWB based term extraction method on four domain specific patent corpora, including *molecular biology (C12N)*, *metallurgy(C22C)*, *electric engineering(G01C)* and *mechanical engineering(H03M)*. Statistics of the documents in each domain are summarized in Table 1.

Table 1. Description of corpora

Corpus	Domain	No. of words	No. of documents
Corpus$_{C12N}$	molecular biology	1,880,739	11,496
Corpus$_{C22C}$	metallurgy	1,915,066	12,226
Corpus$_{G01C}$	mechanical engine.	1,741,820	9,462
Corpus$_{H03M}$	electric engine.	935,059	5,492

It is difficult to collect a complete list of domain terms which is necessary for evaluation. Inspired by the pooling technique used in Information Retrieval (IR), we semi-automatically construct a lexicon for each domain. For each domain, every baseline system and our system submits one term list. The five baseline systems used will be introduced in the next subsections. We take the top 500 terms from every system to form the pool for that domain. Then from the pool, four graduate students majoring in the corresponding domain manually pick out the correct ones to form a pseudo-lexicon for each domain. The sizes of the pseudo-lexicons are 1473, 1380, 1509 and 1370 for molecular biology, metallurgy, mechanical engineering and electric engineering. Then, we compare system performances with the popular IR evaluation measures - precision at 6 different cut-off values P@*n*, where *n*=50, 100, 200, 300, 400, 500.

4.2 Comparison with Other Topic Models

The i-SWB model is the key of our ATE system and we compare it with two commonly used topic models, i.e. SWB (**Baseline 1**) and LDA (**Baseline 2**). Except the construction of topic model, the whole process of term extraction is the same as introduced in Section 3. Fig. 4 below illustrates the P@n values of 3 different topic models on four domains. The blue bars represent the P@n values of our system, the red bars represent the SWB model, and the green bars represent the LDA model. From the figure, we can see that our model significantly outperforms the other two models. Taking the P@50 measure for example, the relative improvement of i-SWB over SWB in the domains of *molecular biology*, *metallurgy*, *mechanical engineering* and *electric engineering* reaches respectively 12.2%, 11.9%, 9.8% and 11.4% respectively.

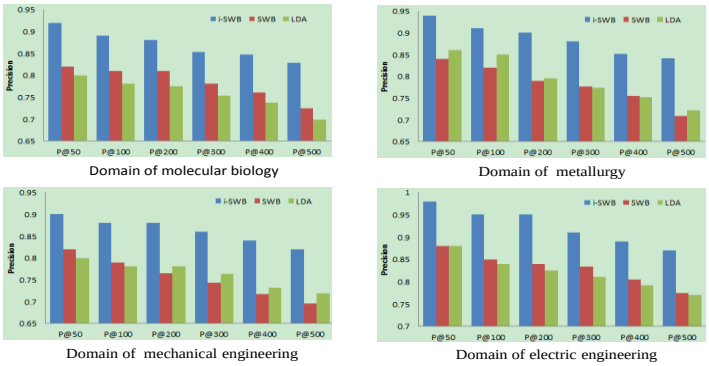

Figure 4. Comparison of different topic models.

When further analyzing the terms that are wrongly identified by each system, we find that if SWB is applied, the word "*signal*" reflecting some general topic is assigned to some document-specific topics due to its high frequency in some documents and then the phrases such as "original signal" and "resulting signal" are wrongly identified as terms. On the other hand, if LDA is used, some general words such as "*purpose*" and "*problem*" which occur frequently in the corpus are assigned to specific topics and the phrases of "*signal purpose*" and "*problem solving*" are wrongly identified as terms. These kinds of problems can be overcome in i-SWB. However, the terms identified by i-SWB that are formed by sequences of several typical words are not always the real terms. For example, "*signal*" and "*component*" are both correctly assigned to general topics with higher probability but "*signal component*" is not a real term. To solve this problem, semantic representation of terms is worth exploring.

4.3 Comparison with Online ATE Service

We also compare our system with three state-of-the-art ATE systems. We use two online free systems - TerMine [2] and TermoStat[3], as **Baseline 3** and **Baseline 4**. TerMine uses the C-value/ NC-value to compute *termhood*, while TermoStat is based on a statistical test and the target items are highly specific to the domain corpus being analyzed. In addition, we implement a baseline system (**Baseline 5**) which adopts TF-IDF in *termhood* computation. Fig. 5 compares our system with the three baseline systems over the P@n values and shows that our system obviously performs better than the other three. This verifies that the semantic analysis of words are helpful to the ATE task. TerMine performs slightly better than TermoStat. But the TF-IDF technique performs unstably: better than TerMine in the metallurgy domain, worse than TerMine and TermoStat in the domain of mechanical engineering. This may suggest that the performance of TF-IDF heavily depends on the corpus quality.

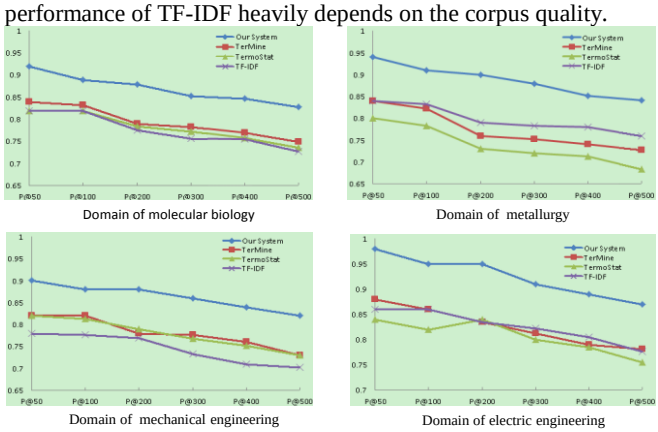

Figure 5. Comparison with other ATE approaches.

5. CONCLUSION

In this paper, we argue that the *termhood* measure should take consideration of semantic information. To cater to this idea, we design a novel topic model i-SWB to map the domain corpus into the latent semantic space, which includes general topics, background topic and documents-specific topic. Based on i-SWB, we implement our ATE system and evaluate it on four domains (i.e. *molecular biology*, *metallurgy*, *mechanical engineering*, and *electric engineering*). The experimental results show that our approach outperforms the state-of-the-art ATE approaches.

6. ACKNOWLEDGMENTS

This research has been supported by NSFC grants (No. 61273278 and 61272291), National Key Technology R&D Program (No:2011BAH1B0403), National 863 Program (No. 2012AA011101) and National Social Science Foundation (No: 12&ZD227). We also thank the anonymous reviewers for their helpful comments. Corresponding authors: Sujian Li and Baobao Chang.

7. REFERENCES

[1] Blei, D. M., Ng, A. Y., Jordan, M.I. 2003. Latent Dirichlet allocation, The Journal of Machine Learning Research, pp. 993-1022.

[2] Chang J.-S. 2005. Domain specific word extraction from hierarchical web documents: a first step toward building lexicon trees from web corpora. In Proceedings of the 4th SIGHAN Workshop on Chinese Language Learning: 64-71.

[3] Chemudugunta, C., Smyth, P. and Steyvers, M. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In NIPS 19, pp. 241–248.

[4] Hisamitsu, T., Niwa, Y., and Tsujii, J. 2000. A method of measuring term representativeness - baseline method using co-occurrence distribution, COLING 2000, pp.320-326.

[5] Frantzi, K., Ananiadou, S. and Mima, H. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal of Digital Libraries, 3(2): 117-132

[6] Kageura, K. and Umino, B. 1996. Methods of automatic term recognition. Terminology, 3(2).

[2] http://www.nactem.ac.uk/software/termine/

[3] http://idefix.ling.umontreal.ca/~drouinp/termostat_web/