Topic Extraction based on Prior Knowledge obtained from Target Documents

Kayo Tatsukawa and Ichiro Kobayashi
Advanced Sciences, Graduate School of Humanities and Sciences,
Ochanomizu University
2-1-1 Ohtsuka Bunkyo-ku Tokyo, 112-8610 JAPAN
{tatsukawa.kayo, koba}@is.ocha.ac.jp

Abstract

This paper investigates the relation between prior knowledge and latent topic classification. There are many cases where the topic classification done by Latent Dirichlet Allocation results in the different classification that humans expect. To improve this problem, several studies using Dirichlet Forest prior instead of Dirichlet distribution have been studied in order to provide constraints on words so as they are classified into the same or not the same topics. However, in many cases, the prior knowledge is constructed from a subjective view of humans, but is not constructed based on the properties of target documents. In this study, we construct prior knowledge based on the words extracted from target documents and provide it as constraints for topic classification. We discuss the result of topic classification with the constraints.

1 Introduction

We have recently faced situations in which we have to deal with a huge amount of text resources. To deal with these text resources, unlike studies to analyze the surface information of the resources, but a lot of studies to analyze latent semantics by means of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been being studied. When extracting latent topics by means of LDA, there are many cases where the words naturally expected to be in the same topic are classified into different topics. To deal with this problem, several studies to provide a constraint for words to be in the same topic have been studied. Andrzejewski (Andrzejewski et al., 2009) has proposed a method to provide a constraint for topic clustering as prior knowledge consisting of the words, which should be in the same topic by applying Dirichlet Forest Prior as word probability distribution instead of Dirichlet distribution. However, in many cases, the prior knowledge is constructed from a subjective view of humans but is not automatically constructed based on the properties of target documents. In this study, we extract the words, which will be prior knowledge for extracting topics, from target documents, and provide it as a constraint for topic clustering, and then discuss the result of topic clustering with constraints on the words.

2 Related studies

Many studies to incorporate prior knowledge into topic models to raise the accuracy of topic clustering, introducing the techniques of semi-supervised learning(Andrzejewski et al., 2007; Andrzejewski et al., 2009; Andrzejewski and Zhu, 2009).

Andrzejewski (Andrzejewski et al., 2009) has incorporated a constraint on words into topic clustering by using Dirichlet Forest Prior instead of Dirichlet distribution. They have introduced 'Must-links' and 'Cannot-links', referring to the techniques of semi-supervised learning. 'Must-links' is a constraint that two words with similar probability distribution should be in the same topic. 'Cannot-links' is a constraint that two words with different probability distribution for all topics should be separated into different topics. Hu (Hu et al., 2011) has proposed a method which repeatedly extracts latent topics through the interaction with humans — constraints are added interactively by humans. In addi-

Proceedings of the 2012 Student Research Workshop, pages 31–36,

Jeju, Republic of Korea, 8-14 July 2012.

tion, Kobayashi (Kobayashi et al., 2011) has made it possible to use logical operation to combine the constraints, 'Must-links' and 'Cannot-links', in constructing prior knowledge. By this, they have proposed a method which can add new constraints constructed by logical operation of various constraints, and extract topics based on the constraints. In general, as for clustering with constraints, it is often that the constraints are given by humans. However, there are many cases where the constraints constructed by humans are arbitrary, in addition, it is laborious to construct prior knowledge for each target document. In this context, Kaji (Kaji et al., 2007) extracted synonyms from corpus by using vocabulary syntactic patterns and constructed prior knowledge for word clustering based on the synonyms. However, the method Kaji proposed obtains prior knowledge by learning approximately 1 billion corpus. So, it also costs much to construct the knowledge, furthermore, the obtained knowledge might be constraints for general purposes, but not for target documents. So, the constructed knowledge might not be appropriate for the target documents.

Considering these things, in this study, we use Dirichlet Forest Prior for word probability distribution and extract latent topics by the prior knowledge obtained from target documents, without using any big corpus. Then we will discuss how our method improves the accuracy of topic extraction.

3 Topic extraction by prior knowledge

3.1 Dirichlet Forest LDA

We use Dirichlet Forest prior (DF) as word probability distribution instead of Dirichlet distribution to reflect constraints on latent topic clustering. DF is hierarchical Dirichlet distribution and it uses α for topic distribution and β for word probability distribution as the hyper-parameters of Dirichlet distribution just like the conventional LDA. In addition, we use η which reflects the strength of given constraints on word occurrence distribution. In Dirichlet Forest, each leave has occurrence probability for each word and the sum of occurrence probability for all words becomes 1. In the process of generating a document with LDA using DF(LDA-DF), we firstly get a multinomial distribution θ with a hyper-parameter α, and then according to this multinomial distribution, a topic Z is selected. Secondly, we get a multinomial distribution ϕ with a hyper-parameter β, and then under the topic Z selected at θ, a word or a constraint is selected. If a word is selected, it is used directly to generate a document and if a constraint is selected, a word is selected according to a multinomial distribution π with hyper-parameter η.

Here, let d_i denote the documents which contain the i-th word w_i and z_i denote the topic which assigns on w_i. Using these parameters, LDA-DF is represented with the below equations.

$$\theta_{d_i} \sim \mathrm{Dirichlet}(\alpha) \quad (1)$$

$$z_i|\theta_{d_i} \sim \mathrm{Multinomial}(\theta_{d_i}) \quad (2)$$

$$q \sim \mathrm{DirichletForest}(\beta, \eta) \quad (3)$$

$$\phi_{z_i} \sim \mathrm{DirichletTree(q)} \quad (4)$$

$$w_i|z_i, \phi_{z_i} \sim \mathrm{Multinomial}(\phi_{z_i}) \quad (5)$$

3.2 Construction of prior knowledge

Newman (Newman et al., 2010) discusses various evaluation indices about the topic coherence. In this study, we choose Point-wise Mutual Information(PMI) as an index to measure topic coherence, and then estimate how much each obtained cluster increases topic coherence in itself. The reason why we choose PMI to measure topic coherence is based on the assumption that a topic is represented by the words with close relationship.

To construct prior knowledge, it is necessary to select words regarded as representatives of a topic. In this study, we assume that the words regarded as representatives of a topic ('important words', hereafter) frequently appear in all documents or have many co-occurrence relations with a lot of other words. We select important words by following the two basic ideas shown below.

(i) Important words based on frequency

In the case of dealing with multiple documents about the same topic, the words which frequently appear in all documents are regarded as necessary words to represent the contents of the documents. So, we regard such words as important words.

(ii) Important words based on co-occurrence

In this study, we construct prior knowledge as we suppose that a pair of words with high PMI value

should be classified into the same topic. So, we regard the words, which have many co-occurrence relations with other words, as important words.

The prior knowledge is constructed by the following process.

step.1 Important words based on frequency or co-occurrence are selected.

step.2 Important words obtained at step.1 are classified into some groups based on co-occurrence relation. At this time, we use PMI as index to measure co-occurrence relation between words, and unite important words, which have higher PMI than the predefined threshold value, into a group.

step.3 Prior knowledge, i.e., the group obtained at step2, is constructed based on the words with high PMI values, therefore, the words which have high PMI value with the words in the group obtained at step.2 are further selected and added to the group, if necessary. Depending on the number of words added to the group, prior knowledge will be changed. So, we experiment to investigate the influence of the number of added words, changing the number of the words from 1 to 4. The detail about the experiment is mentioned in section 4.

4 Experiment

4.1 Experimental settings

As the documents for the experiment to extract topics, we used news articles about the same incident. The news articles we used are ABC News in USA, BBC News in UK, CTV News in Canada, which are published by main newspaper companies and TV companies in English-speaking countries.

We used the following 4 articles for the experiment: 10 articles about 'Press conference about the convergence of atomic power plant disaster by Japanese prime minister, 2011/12/16' consist of 212 documents and 853 terms; 24 articles about 'Grounding of pomp passenger ferry in Italy, 2012/1/16' consist of 967 documents and 2267 terms; 25 articles about 'Protest from Wikipedia to Stop Online Piracy Act (SOPA), 2012/1/16' consist of 700 documents and 1823 terms; 18 articles about 'Resignation of co-founder Yahoo!, 2012/1/16' consist of 553 documents and 1113 terms.

In the experiment, we used $\alpha = 0.1$, $\beta = 0.1$, $\eta = 100$ as hyper-parameters for LDA-DF and Collapsed Gibbs Sampling (Griffiths et al., 2004) for the presumption of probability distribution with 50 iteration times.

Although we could set the number of topics so as it fitted target documents by means of perplexity, since we aim to evaluate adequacy of grouping of words, adequacy of topic clustering in other words, in the same condition, so we conducted an experiment, setting the number of topics as 10 for all the target articles.

In Hu's study (Hu et al., 2011), in response to given words as constraints, they re-presumed latent topics by canceling a part of the topics already assigned to words by the topic model prior to addition of new constraints. They suggested 4 ways of selecting words to cancel a part of topics, and reported that in the 4 ways they got good results when new prior knowledge is added, topic assignment for all the words of the documents which include the words in the prior knowledge is once canceled and then applied again. Therefore, we also cancel the topics assigned to words in the same way of theirs.

We calculate the value of perplexity of topic distribution and compare the stability of a model between before and after giving constraints. we calculate perplexity with equation (6). Here, N is the number of all words in the target documents, w_{mn} is the n-th word in the m-th document; θ is occurrence probability of topic for the documents, and ϕ is occurrence probability of the words for every topic.

$$Perplexity(\boldsymbol{w}) = exp(-\frac{1}{N}\sum_{mn} log(\sum_{z} \theta_{mz}\phi_{z w_{mn}})) \quad (6)$$

4.2 Experiment result

Table 1 shows the groups of important words based on frequency and co-occurrence, and the words with high PMI score to the important words which are candidates to be added to the prior knowledge. We take up the article about 'Press conference about the convergence of atomic power plant disaster by Japanese prime minister' and explain how to interpret Table 1.

Looking at the intersection between the row of frequency and the column of 'Atomic plant'

Table 1: Groups of important words based on frequency or co-occurrence, and added words

Types/Articles		Atomic plant	Grounding of ferry	Protest to SOPA	Yahoo! co-founder
Frequency	grouping words	{prime,minister,reactor, fukushima}, {power,tokyo},{cold}, {nuclear},{plant}, {shutdown}.	{costa},{passenger}, {people}	{wikipedia},{online}, {piracy},{internet}	{yang},{board}, {yahoo},{company}, {thompson}
	added words	yoshihiko,electric,reached, noda,march,state	appears,unaccounted, friday	wale,stop,protect,free	bostock,position, chairman,struggling,scott
Co-occurrence	grouping words	{cooling,contaminated, water},{site},{year} {stable,state,response}, {worst,disaster}	{disaster,caused,sea}, {aground,ran},{gash}, {authority,safety}, {television},{evacuation}	{medium,industry,group, tech,information,popular}, {big},{legislation}, {service},{community}	{private,pursuing,deal, shareholder,asian}, {began},{leaving}, {resignation},{chief}, {medium}
	added words	ton,liquid,end,tank, chernobyl	technical,late,side, trained,human,survivor	social,web,proposed, provider,wale	large,university, struggling,thompson, scott,trading

in Table 1, the extracted important words were united to one group depending on the value of PMI, and then we obtained the following 6 groups: {prime, minister, reactor, fukushima},{power, tokyo},{cold},{nuclear},{plant},{shutdown}. After that, we added some words with high PMI value to each group to achieve the construction of prior knowledge. The words expected to be added to the groups are shown at the next row of grouping words. In fact, depending on the number of the words added to the groups of important words, prior knowledge will be changed and the result of topic clustering will also be changed. Furthermore, depending on the number of given constraints, the result of topic clustering will also be changed. Therefore, we examine how accuracy of topic clustering changes by means of perplexity as its index, increasing the number of given constraints one by one from the initial condition, i.e., without any constraint.

Here, we think that the values of PMI and perplexity will be changed by the combination of prior knowledge, however, in this study, we gave the constraints in the order of a group with higher PMI value.

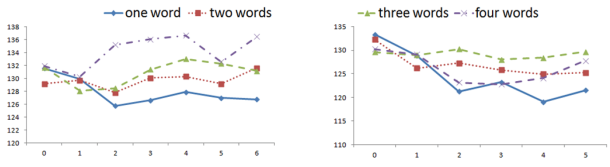

Figure 1: 'Convergence of atomic plant disaster'

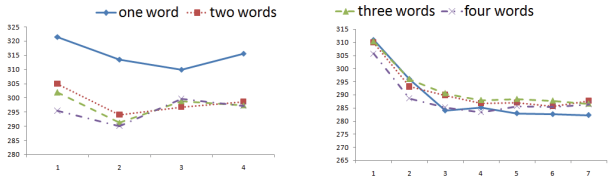

Figure 2: 'Grounding of pomp passenger ferry in Italy'

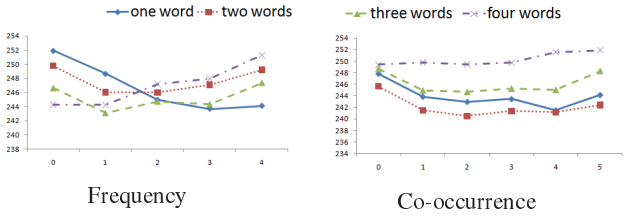

Figure 3: 'Protest of Wikipedia to SOPA'

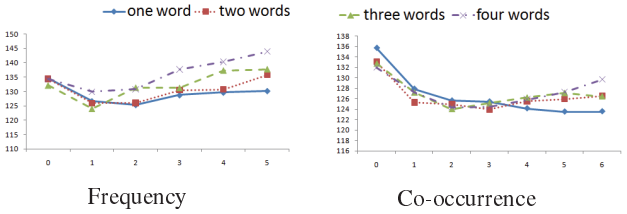

Figure 4: 'Resignation of Yahoo! co-founder'

Table 2: Top 10 representative words for topics extracted from the article 'Convergence of atomic plant disaster'

topic	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
LDA	water plant contaminated remains decade expert facility cool point problem	plant return home government zone resident remain evacuation mile doe	nuclear task told crisis meeting force disaster nod situation response	nuclear tsunami crisis march plant earthquake announcement fukushima meltdown month	cold shutdown reactor plant fukushima condition government stable power reached	accident nuclear disaster chernobyl university term country part engineering professor	tokyo electric power leak time week knocked huge bring official	year radiation plant area level government expected official start boundary	cooling minister reactor prime noda system degree yoshihiko nuclear temperature	reactor fuel tepco temperature rod melted spent inside damaged damage
topic	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
LDA-DF with constraints	facility waste tepco doe contaminated water including cooling earlier sea	plant home return mile zone government area resident official remain	nuclear noda task power force meeting japan set measure declared	tsunami plant march earthquake reactor crisis meltdown system daiichi people	cold shutdown plant reactor condition fukushima government reached friday radiation	nuclear cleanup university country significant disaster chernobyl worst term engineering	tokyo electric government power official told week bring company tepco	plant radiation level year end decade government decommission expert accident	nuclear prime minister degree announcement yoshihiko noda milestone mark news	reactor fuel temperature tepco rod fukushima melted cool spent inside

4.3 Discussions

We show the changes of perplexity in Figure 1, 2, 3, and 4 when increasing the number of constraints based on frequency and co-occurrence. In the Figures, the horizontal axis indicates the number of pieces of prior knowledge, and the vertical axis indicates the value of perplexity. Looking at these Figures, we see that the case of providing constraints based on co-occurrence decreases perplexity as the number of constraints increases, and the topic model becomes more stable than the case without constraint. Furthermore, we see that perplexity of each graph of co-occurrence can be decreased if providing one or two additional words with high value of PMI as a part of prior knowledge, and also that perplexity becomes stable when approximately 3 constraints are provided. From these observations, we think that we do not have to provide so many constraints to get good topic clustering.

On the other hand, unlike the case of providing constraints based on co-occurrence, we cannot get a general view for the case of providing constraints based on frequency from the results.

The reason why we could get good results when providing constraints based on co-occurrence information is that we constructed the prior knowledge which simultaneously reflects both 'Must-Links' and 'Cannot-Links' used as prior knowledge in (Andrzejewski et al., 2009), because PMI represents the co-occurrence relation of words in a sentence, so we think that it could divide the words should be included or should not be included in a topic.

On the other hand, we also see the case where perplexity gets increased even if selecting important words based on co-occurrence. Looking at the case of providing four additional words to prior knowledge in the graph of co-occurrence, especially in Figure 2,3,and 4, we see that perplexity increases as the number of pieces prior knowledge increases. We think the reason for this is because we added words to prior knowledge in the order of high PMI value, so the fourth word should not have had high PMI value, therefore, topic clusters became unstable.

Table 2 shows the result of topic classification of the article about 'Press conference of the convergence of atomic power plant disaster by Japanese prime minister.' We added the following constraints as prior knowledge: {worst, disaster, chernobyl},{cooling, contaminated, water, ton}, and {year, end} which is constructed based on co-occurrence information in the objective article. The upper row of Table 2 is the result of the conventional LDA without any constraint and the lower row is that of LDA-DF with constraints.

We see from Table 2 that the words consisting of

prior knowledge are split into two topics at the upper row, whereas, they are classified in the same topic, i.e., topic 0,5,and 7 at the lower row. We see that topic clustering with the constraints has been well achieved.

5 Conclusion

The conventional LDA sometimes results in topic classification different from what humans expect. To improve this, several studies providing constraints for topic clustering have been studied, referring to the techniques of semi-supervised learning.

In this study, we have constructed prior knowledge, which becomes constraints for topic clustering, with target documents which topics are extracted, unlike the studies to construct the knowledge with huge corpus. The prior knowledge will be constructed as a collection of the words expected to be representative of a topic. Based on this, we have introduced two ways to construct the knowledge: one is to select important words based on frequency and the other is to select words based on co-occurrence from target documents. We have compared the results of topic clustering by giving the two types of prior knowledge, and then recognized that the result of topic clustering based on the prior knowledge constructed based on co-occurrence is better than that by the prior knowledge constructed based on frequency. Furthermore, we have also investigated how much prior knowledge should be given as constraints for good topic clustering, and then obtained a result that good clustering is achieved even with a few pieces of prior knowledge, if the prior knowledge is constructed based on word co-occurrence. However, we have also observed several cases where this result cannot be correct. We need more investigation about this, revising the way of constructing prior knowledge. For future work, we will investigate another possibility to construct prior knowledge, and will apply our proposed method to various kinds of many documents.

References

David M. Blei and Andrew Y. Ng and Michael I. Jordan and John Lafferty. 2003. *Latent dirichlet allocation*, Journal of Machine Learning Research,

Hayato Kobayashi and Hiromi Wakaki and Tomohiro Yamasaki and Masaru Suzuki 2011. *Topic Models with Logical Constraints on Words*, Proc. of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing,

Andrzejewski, Anne Mulhern, Ben Liblit, and Xiaojin Zhu, 2007. *Statistical Debugging Using Latent Topic Models*, Proceedings of the 18th European Conference on Machine Learning (ECML2007), pp. 6–17, Springer-Verlag.

Andrzejewski, David and Zhu, Xiaojin and Craven, Mark, 2009. *Incorporating domain knowledge into topic modeling via Dirichlet Forest priors*, Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09, pp. 25–32, Montreal, Quebec, Canada.

Andrzejewski, David and Zhu, Xiaojin, 2009. *Dirichlet Allocation with Topic-in-Set Knowledge* Proceedings of NAACL-HLT2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pp. 43–48.

Hu, Yuening and Boyd-Graber, Jordan and Satinoff, Brianna, 2011. *Interactive topic modeling*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pp.248–257, Portland, Oregon, USA.

Nobuhiro Kaji and Masaru Kitsuregawa, 2007. *Constrained distributional clustering of words using lexico-syntactic patterns (in Japanese)*, SIG-KBS, 79, pp.61-66, 2007-12-03.

Newman, David and Lau, Jey Han and Grieser, Karl and Baldwin, Timothy, 2010. *Automatic evaluation of topic coherence*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108,Los Angeles, California.

S.Y.Dennis III. 1991. *On the Hyper-Dirichlet Type 1 and Hyper-Liouville distributions.*, Communications in Statics – Theory and Methods 20(12):pp.4069-4081.

Minka, T.P. 1999. *The Dirichlet-tree distribution*, (Technical Report) http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirtree.pdf

Kristina Toutanova and Mark Johnson. 2008. *A Baysian LDA-based model for semi-supervised part-of-speech tagging.*, In Advances in Neural Information Processing Systems 20, pp.1521-1528, MIT Press.

Thomas L.Griffiths and Mark Steyvers. 2004. *Finding scientific topics*, Proceedings of the National Academy of Sciences, Vol.101,No.Suppl 1. pp.5228-5235.