News Thread Extraction Based on Topical N-Gram Model with a Background Distribution

Zehua Yan and Fang Li

Department of Computer Science and Engineering,
Shanghai Jiao Tong University
{yanzehua,fli}@sjtu.edu.cn
http://lt-lab.sjtu.edu.cn

Abstract. Automatic thread extraction for news events can help people know different aspects of a news event. In this paper, we present a method of extraction using a topical N-gram model with a background distribution (TNB). Unlike most topic models, such as Latent Dirichlet Allocation (LDA), which relies on the bag-of-words assumption, our model treats words in their textual order. Each news report is represented as a combination of a background distribution over the corpus and a mixture distribution over hidden news threads. Thus our model can model "presidential election" of different years as a background phrase and "Obama wins" as a thread for event "2008 USA presidential election". We apply our method on two different corpora. Evaluation based on human judgment shows that the model can generate meaningful and interpretable threads from a news corpus.

Keywords: news thread, LDA, N-gram, background distribution.

1 Introduction

News events happen every day in the real world, and news reports describe different aspects of the events. For example, when an earthquake occurs, news reports will report the damage caused, the actions taken by the government, the aid from the international world, and other things related to the earthquake. News threads represent these different aspects of an event.

Topic models, such as Latent Dirichlet Allocation (LDA) [1] can extract latent topics from a large corpus based on the bag-of-words assumption. Actually news reports are sets of semantic units represented by words or phrases. N-gram phrases are meaningful to represent these semantic units. For example, "Bush Government" and "Security Council" in table 1 are two news threads for the "Iran nuclear program" event. They capture two aspects of the meaning of the event reports. Our task is to automatically extract news threads from news reports.

Reports of a news event or a topic discuss the same event or the same topic and share some common words. Based on the analysis of LDA results, we find that such common words represent the background of the event. We then assume each

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part II, LNCS 7063, pp. 416–424, 2011.

news report is represented by a combination of (a) a background distribution over the corpus, (b) a mixture distribution over hidden news threads.

In this paper, we use a topical n-gram model with a background distribution (TNB) to extract news threads from a news event corpus. It is an extension of the LDA model with word order and a background distribution. In the following, our model will be introduced, then experiments described and results given.

Table 1. Threads and news titles for news event "Iran nuclear program"

Event corpus	Thread	News report titles
Iran Nuclear Program	the Security Council	Options for the Security Council
		Iran ends cooperation with IAEA
		Iran likely to face Security Council
	the Bush government	Rice: Iran can have nuclear energy, not arms
		Bush plans strike on Iran's nuclear sites
		Iran Details Nuclear Ambitions

2 Related Work

In [2]'s work, news event threading is defined as the process of recognizing events and their dependencies. They proposed an event model to capture the rich structure of events and their dependencies in a news topic. Features such as temporal locality of stories and time-ordering are used to capture events.

[3] proposed a probabilistic model that accounts for both general and specific aspects of documents. The model extends LDA by introducing a specific aspect distribution and a background distribution. In this paper, each document is represented as a combination of (a) a background distribution over common words, (b) a mixture distribution over general topics, and (c) a distribution over words that are treated as being specific to the documents. The model has been applied in information retrieval and showed that it can match documents both at a general level and at specific word level. Similarly, [4] proposed an entity-aspect model with a background distribution; the model can automatically generate summary templates from given collections of summary articles.

Word order and phrases are often critical to capture the latent meaning of text. Much work has been done on probabilistic generation models with word order influence. [5] develops a bigram topic model on the basis of a hierarchical Dirichlet language model [6], by incorporating the concept of topic into bigrams. In this model, word choice is always affected by the previous word.

[7] proposed an LDA collocation model (LDACOL). Words can be generated from the original topic distribution or the distribution in relation to the previous word. A new bigram status variable is used to indicate whether to generate a bigram or a unigram. It is more realistic than the bigram topic model which always generates bigrams. However, in the LDA Collocation model, bigrams do not have topics because the second term of a bigram is generated from a distribution conditioned on its previous word only.

Further, [8] extended LDACOL by changing the distribution of previous words into a compound distribution of previous word and topic. In this model, a word has the option to inherit a topic assignment from its previous word if they form a bigram phrase. Whether to form a bigram for two consecutive word tokens depends on their co-occurrence frequency and nearby context.

3 Our Methods

3.1 Motivation

We analyze different news reports, and find that there are three kinds of words in a news report: background words (B), thread words (T) and stop words (S). Background words describe the background of the event. They are shared by reports in the same corpus. Thread words illustrate different aspects of an event. Stops words are meaningless and appear frequently across different corpora.

For example, there are two sentences from a news report of "US presidential election" in table 2. The first sentence talks about "immigration policy" and the second discusses "healthcare". Stop words are labeled with "S" such as "as" and "the". Background words are "presidential" and "election" which appear in both sentences and are labeled with "B". Other words are thread words that are specifically associated with different aspects of the event, such as "immigration" and "healthcare".

Table 2. Two sentences from "US presidential election"

As/S we/S approach the/S 2008 Presidential/B election/B,/S both/S John/B McCain/B and/S Barack/B Obama/B are/S sharpening/T their/S perspectives/B on/S immigration/T policy/B./S
After/S the/S economy/T ,/S US/B healthcare/T is/S the/S biggest/T domestic/T issue/T influencing/B voters/B in/S the/S US/B presidential/B election/B ./S

Also, we note that adjacent words can form a meaningful phrase and provide a clearer meaning, for example, "presidential election" and "domestic issue". Based on the analysis, there are four possible combinations as follows:

1. B+B: Presidential/B election/B
2. B+T: US/B healthcare/T
3. T+B: immigration/T policy/B
4. T+T: domestic/T issue/T

There is no doubt that "B+B" is a background phrase, and the "T+T" is a thread phrase. Both "B+T" and "T+B" are regarded as thread phrases because the phrase contains a thread word. For example, immigration is a thread word and policy is a background word; the phrase "immigration policy" identifies a type of "policy", and should be viewed as a thread phrase.

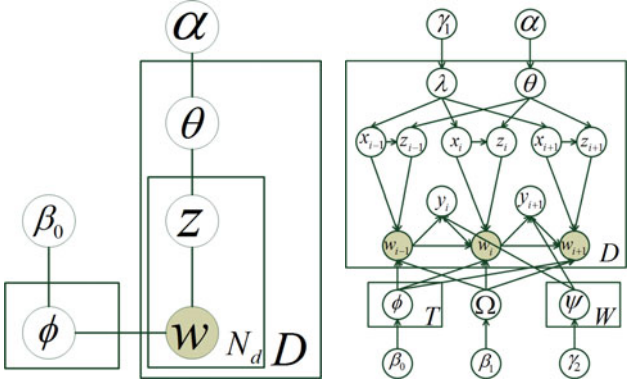
3.2 Topical N-Gram Model with Background Distribution

We now propose our topical n-gram model with a background distribution (TNB) for news reports. Notation used in this paper is listed in table 3. Stop words are identified and removed using a stop word list.

In our model, each news report is represented as a combination of two kinds of multinomial word distribution:

(a) There is a background word distribution Ω with Dirichlet prior parameter β_1, which generates common words across different threads. (b) There are T thread word distributions $\phi_t (1 < t < T)$ with Dirichlet prior parameter β_0. A hidden bigram variable x_i is used to indicate whether a word is generated from the background word distribution or the thread word distribution.

A hidden bigram variable y_i is introduced to indicate whether word w_i can form a phrase with its previous word w_{i-1} or not. Unlike [8], we assume phrase generation is only affected by the the previous word.



(a) LDA (b) TNB

Fig. 1. Graphical model for LDA and TNB

Figure 1 shows graphical models of LDA and TNB. For each word w_i, LDA first draws a topic z_i from the document-topic distribution $p(z|\theta_d)$ and then draws the word from the topic-word distribution $p(w_i|\phi_{z_i})$. TNB has a similar general structure to the LDA model but with additional machinery to identify word w_i's category (background or thread word) and whether it can form a phrase with the previous word w_{i-1}.

For each word w_i, we first sample variable y_i. If $y_i = 0$, w_i is not influenced by w_{i-1}. If $y_i = 1$, w_{i-1} and w_i can form a phrase. As analyzed before, phrases have four possible combinations. There are two situations when $y_i = 1$:

1. if $w_{i-1} \in z_t$, w_i draws either from the thread z_t or the background distribution.

2. if w_{i-1} is a background word, w_i draws from any threads or the background distribution.

Table 3. Notation used in this paper

SYMBOL	DESCRIPTION	SYMBOL	DESCRIPTION
α	Dirichlet prior of θ	β_0	Dirichlet prior of ϕ
β_1	Dirichlet prior of Ω	γ_1	Dirichlet prior of λ
γ_2	Dirichlet prior of σ	T	number of threads
D	number of documents	W	number of unique words
$w_i^{(d)}$	the i^{th} word in document d	$z_i^{(d)}$	the thread associated with i^{th} word in the document d
$y_i^{(d)}$	the bigram status between the $(i-1)^{th}$ word and i^{th} word in the document d	$x_i(d)$	the bigram status indicate the i^{th} word is a background word or topic word
$\theta^{(d)}$	the multinomial distribution of topics w.r.t the document d	ϕ_z	the multinomial distribution of words w.r.t the topic z
Ω	the multinomial distribution of words w.r.t the background	ψ_i	the Bernoulli distribution of status variable $y_i(d)$
λ_i	the Bernoulli distribution of status variable $x_i(d)$		

Second, we sample variable x_i. If $x_i = 1$, w_i is a background word, it is generated from $Multi(\Omega)$. Else it is generated in the same way as LDA.

3.3 Inference

For this model, exact inference over hidden variables is intractable due to the large number of variables and parameters. There are several approximate inference techniques which can be used to solve this problem, such as variational methods [9], Gibbs sampling [10] and expectation propagation [11]. As [12] showed that phrase assignment can be sampled efficiently by Gibbs sampling, Gibbs sampling is adopted for approximate inference in our work.

The conditional probability of w_i given a document d_j can be written as:

$$\begin{aligned} p(w_i|d_j) = (&p(x_i = 0|d_j) \textstyle\sum_{t=1}^{T} p(w_i|z_i = t, d) \\ &+ p(x_i = 1|d_j) p'(w)) \times p(w_i|y_i, w_{i-1}) \end{aligned} \tag{1}$$

where $p(w_i|z_i = t, d)$ is the thread word distribution and $p'(w)$ is the background word distribution. $p(w_i|y_i, w_{i-1})$ describe the $w'_{i-1}s$influence over w_i.

In Figure 1(b), if $y_i = 0$, the w_i will not be influenced by w_{i-1} and will be generated from the background distribution and thread distribution. Gibbs sampling equations are derived as follows:

$$\begin{aligned} &p(x_i = 0, y_i = 0, z_i = t|w, x_{-i}, z_{-i}, \alpha, \beta_0, \gamma_1, \gamma_2) \propto \\ &\frac{N_{d0,-i}+\gamma_1}{N_{d,-i}+2\gamma_1} \times \frac{C^{TD}_{td,-i}+\alpha}{\sum_{t'} C^{TD}_{t'd,-i}+T\alpha} \times \frac{C^{WT}_{wt,-i}+\beta_0}{\sum_{w'} C^{WT}_{w't,-i}+T\beta_0} \times \frac{N_0^{w_{i-1}}+\gamma_2}{N_{w_i-1}+2\gamma_2} \end{aligned} \tag{2}$$

$$\begin{aligned} &p(x_i = 1, y_i = 0|w, x_{-i}, z_{-i}, \beta_1, \gamma_1, \gamma_2) \propto \\ &\frac{N_{d1,-i}+\gamma_1}{N_{d,-i}+2\gamma_1} \times \frac{C^{W}_{w,-i}+\beta_1}{\sum_{w'} C^{W}_{w',-i}+T\beta_1} \times \frac{N_0^{w_{i-1}}+\gamma_2}{N_{w_i-1}+2\gamma_2} \end{aligned} \tag{3}$$

If $y_i = 1$, the w_i can form a phrase with w_{i-1}.

$$p(x_i = 0, y_i = 1, z_i = t | w_{i-1}, z_{i-1} = t, \alpha, \beta_0, \gamma_1, \gamma_2) \propto \frac{N_{d0,-i}+\gamma_1}{N_{d,-i}+2\gamma_1} \times \frac{C^{WT}_{wt,-i}+\beta_0}{\sum_{w'} C^{WT}_{w't,-i}+T\beta_0} \times \frac{N_1^{w_{i-1}}+\gamma_2}{N_{w_i-1}+2\gamma_2} \tag{4}$$

$$p(x_i = 1, y_i = 1 | w_{i-1}, z_{i-1} = t, \alpha, \beta_1, \gamma_1, \gamma_2) \propto \frac{N_{d1,-i}+\gamma_1}{N_{d,-i}+2\gamma_1} \frac{C^{W}_{w,-i}+\beta_1}{\sum_{w'} C^{W}_{w',-i}+T\beta_1} \times \frac{N_1^{w_{i-1}}+\gamma_2}{N_{w_i-1}+2\gamma_2} \tag{5}$$

where the subscript $-i$ stands for the count when word i is removed. N_d is the number of words in document d. N_{d0} stands for the number of thread words in document d, and N_{d1} is the number of background words in document d. $N_{w_{i-1}}$ is the number of words w_{i-1}. $N_0^{w_{i-1}}$ and $N_1^{w_{i-1}}$ is the number of words w_{i-1} which have been drawn from as a unigram or as a part of phrase. C^{WT}_{wt}, C^{W}_{w} are the number of times a word is assigned to a thread t, or to a background distribution respectively.

4 Experiments

4.1 Experimental Settings

Two corpora are used in the experiments. The Chinese news corpus is an event based corpus, which contains 68 event sub-corpora, such as "2007 Nobel prize". The number of news reports in a sub-corpus varies from 100 to 420. Another corpus is the Reuters-21578 financial news corpus. We select five sub corpora from it, they are: "crude", "grain", "interest", "money-fx" and "trade". Each of them contains more than 300 reports which describe many events.

Experiments are run on both corpora with different numbers of threads. The experiments are run with 500 iterations for each case. And we set $\alpha = 50/T$ where T is the number of threads, $\beta_0 = 0.1$, $\beta_1 = 0.1$ and $\gamma_1 = 0.5$, $\gamma_2 = 0.5$ by experience.

The LDA result is used as our baseline. The top three words of LDA are compared with the top three phrases generated by TNB on different corpora at different numbers of threads.

4.2 Evaluation Metrics

There is no golden standard for news thread extraction. Only humans can identify and understand news threads for different news events. The top three phrases of TNB and top three words of LDA are evaluated by voluntary judges on a scale of 0 to 1. Report titles are provided as the basis for judging. Score 1 means the phrase or the word represents the meaning of the title well. Score 0 means the word or the phrase does not capture the meaning of the title. Score 0.5 is between them. The precision of news threads are calculated in the following three formula:

$$top{-}1 = \frac{\sum_t^T score_{t1}}{T} \tag{6}$$

$$top-2 = \frac{\sum_t^T max(score_{t1}, score_{t2})}{T} \tag{7}$$

$$top-3 = \frac{\sum_t^T max(score_{t1}, score_{t2}, score_{t3})}{T} \tag{8}$$

where $score_{ti}$ is the score of the i^{th} word in thread t.

4.3 Results and Analysis

Table 4 and 5 shows the precisions of news thread extraction from the Chinese and Ruters corpus with different numbers of threads. As the number of thread increases, the precision decreases. We analyze both corpora. The Chinese corpus is event-based, the number of 5 or 8 matches its semantic meaning hidden in each event corpus. Twenty threads are adequate to the semantic meanings of the Reuters sub-corpora. The hidden semantics of the corpus dominate the precision and final results.

The precision of TNB is much better than LDA. We give two explanations. Table 7 shows both results extracted from the "2007 Nobel Prize" reports. First, the top LDA words do not consider the background influence, common words such as "Nobel" appearing in the top three words. Such words cannot be regarded as thread words to represent different aspects of an event. In TNB, thread-specific words (such as "Peace") can be extracted and form an n-gram phrase with backgroun word to represent the thread more clearly. The second explanation is that a phrase delivers more clear information than a unigram word. For example, "peace" vs. "Nobel Peace Prize". The top three results of TNB for threads related to the Nobel Peace Prize convey two meanings "Nobel Peace Prize" and "Climate change problem", while people need his knowledge to understand the top three words of LDA.

Table 4. Precision on Chinese corpus

Evaluations	Number of thread			
	5	8	10	12
TNB top-1	72.3%	65.4%	61.5%	60.9%
TNB top-2	85.2%	82.4%	77.7%	75.1%
TNB top-3	90.6%	88.3%	82.9%	81.4%
LDA top-1	43.4%	38.3%	31.9%	30.3%
LDA top-2	51.3%	45.5%	37.5%	36.9%
LDA top-3	58.4%	55.1%	46.9%	43.3%

Table 5. Precision on Reuter corpus

Evaluations	Number of thread		
	20	25	30
TNB top-1	55.2%	44.3%	38.3%
TNB top-2	73.2%	61.1%	57.7%
TNB top-3	81.3%	69.4%	66.3%
LDA top-1	32%	29.5%	28.3%
LDA top-2	41.5%	37%	38.4%
LDA top-3	52%	41.5%	40%

Table 6 lists the background words of five sub-corpora of Reuters news. These sub-corpora are not event-based, The background words still catch many features of each category. For example, words like "wheat", "grain" and "agriculture" are easily identified as background words for the category of grain. The word "say" appears as the top background word for all these sub-corpora. The reason is that reports in the Reuters corpus always reference different peoples' opinions, so the word frequency is really high. Therefore "say" is regarded as a background word.

Table 6. Background words for Reuters corpus

trade	crude	grain	interest	money-fx
say	say	say	say	say
trade	oil	wheat	rate	dollar
japan	company	price	bank	rate
japanese	dlrs	grain	market	blah
official	mln	corn	blah	trade

Table 7. LDA and TNB result for threads of "2007 Nobel prize"

Nobel Peace Prize		**Nobel Economics Prize**	
LDA Result			
Peace	0.032	Nobel	0.041
Nobel	0.025	Sweden	0.035
Climate	0.024	economics	0.029
Gore	0.023	announce	0.027
change	0.019	prize	0.021
president	0.016	date	0.015
committee	0.013	winner	0.014
global	0.013	economist	0.013
TNB Background words			
America	0.015	research	0.013
university	0.013	nobel	0.012
gene	0.011	Prize	0.011
TNB Result			
Nobel Peace Prize	0.033	The Royal Swedish Academy	0.056
Climate change problem	0.032	announce Nobel economics prize	0.052
Climate change	0.018	Swedish kronor	0.038

5 Conclusion

In this paper, we present a topical n-gram model with background distribution (TNB) to extract news threads. The TNB model adds background analysis and the word-order feature to standard LDA. Experiments indicate that our model can extract more interpretable threads than LDA from a news corpus. We also find that the number of threads and the event type can influence the precision of news thread extraction. Experiments show that TNB works well not only on an event-based corpus but also on a topic-based corpus. In the future, we plan to develop a dynamic mechanism to decide a suitable number of threads for different news event types to improve the precision of news thread extraction.

Acknowledgements. This research is supported by the Chinese Natural Science Foundation under Grant Numbers 60873134.The authors thank Mr.Sandy Harris for English improvment and other students for human evaluations in the experiments.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Nallapati, R., Feng, A., Peng, F., Allan, J.: Event threading within news topics. In: Proceedings of the Thirteenth ACM International Conference on Information and knowledge Management, pp. 446–453. ACM (2004)
3. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: Advances in Neural Information Processing Systems, pp. 241–242 (2006)
4. Li, P., Jiang, J., Wang, Y.: Generating templates of entity summaries with an entity-aspect model and pattern mining. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 640–649. Association for Computational Linguistics (2010)
5. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984. ACM (2006)
6. MacKay, D.J.C., Peto, L.C.B.: A hierarchical dirichlet language model. Natural language engineering 1(03), 289–308 (1995)
7. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review 114(2), 211 (2007)
8. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Seventh IEEE International Conference on Data Mining ICDM 2007, pp. 697–702. IEEE (2007)
9. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Machine learning 37(2), 183–233 (1999)
10. Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.I.: An introduction to mcmc for machine learning. Machine learning 50(1), 5–43 (2003)
11. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, pp. 352–359. Citeseer (2002)
12. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228 (2004)