Modeling Chinese Documents with Topical Word-Character Models

Wei Hu[1] **Nobuyuki Shimizu**[2]
[1]Department of Computer Science
Shanghai Jiao Tong University
Shanghai, China 200240
{no_bit,hysheng}
@sjtu.edu.cn

Hiroshi Nakagawa[2] **Huanye Sheng**[1]
[2]Information Technology Center
The University of Tokyo
Tokyo, Japan 113-0033
{shimizu, nakagawa}
@r.dl.itc.u-tokyo.ac.jp

Abstract

As Chinese text is written without word boundaries, effectively recognizing Chinese words is like recognizing collocations in English, substituting characters for words and words for collocations. However, existing topical models that involve collocations have a common limitation. Instead of directly assigning a topic to a collocation, they take the topic of a word within the collocation as the topic of the whole collocation. This is unsatisfactory for topical modeling of Chinese documents. Thus, we propose a topical word-character model (TWC), which allows two distinct types of topics: *word topic* and *character topic*. We evaluated TWC both qualitatively and quantitatively to show that it is a powerful and a promising topic model.

1 Introduction

Topic models (Blei et al., 2003; Griffiths & Steyvers 2004, 2007) are a class of statistical models in which documents are expressed as mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a probabilistic procedure for generating documents. To make a new document, we choose a distribution over topics. Then, for each word in this document, we randomly select a topic from the distribution, and draw a word from the topic. Once we have a topic model, we can invert the generating process, inferring the set of topics that was responsible for generating a collection of documents.

Although most topic models treat a document as a bag-of-words, the assumption has obvious shortcomings. Suppose there are many documents about art and musicals in New York. Then, we find a topic represented by words such as "art", "musical", and "New", instead of getting "New York". The bag-of-words assumption makes the topic model split a collocation—a phrase with meaning beyond the individual words—into individual words that have a different meaning. One example of a collocation is the phrase "white house". In politics, it carries a special meaning beyond a house that is white, whereas "yellow house" does not.

While it is reasonable for English, the equivalent bag-of-characters assumption is especially troublesome for modeling Chinese documents, where almost all basic vocabularies are the equivalents of English collocations. In Chinese, some of the most commonly used couple-of-thousand characters are combined to make up a word, and no word boundary is given in the text. Effectively, Chinese words are like collocations in English. The difficulty is that there are overwhelmingly more of them, enough to render a bag-of-character assumption unreasonable for Chinese. Therefore, a topical model for Chinese should be capable of detecting the boundary between two words, as well as assigning a topic to each word.

While topic models for Chinese documents bear some similarity to collocation models in English, existing topical collocation discovery models, such as the LDA (latent Dirichelt allocation) Collocation model (LDACOL) (Griffiths et al., 2007) and the topical N-gram model (TNG) (Wang et al., 2007), do not directly assign a topic to a collocation. These models find the boundaries of phrases and assign a topic to each word. The problem is in the next step—the topic of the collocation is exactly the same as one of the words. This is like saying that the topic of "white

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 345–352
Manchester, August 2008

house" is the same as either that of "white" or "house". We propose a new topical model, the topical word-character model (TWC), which aims to overcome these limitations.

We evaluated the model both quantitatively and qualitatively. For the quantitative analysis, we compared the performance of TWC and TNG using a standard measure *perplexity*. For the qualitative analysis, we evaluated TWC's ability to discover Chinese words and assign topics in comparison with TNG.

The rest of the paper is organized as follows. Section 2 reviews topic models that aim to include collocations explicitly in the model and analyzes their limitations. Section 3 presents our new model TWC. Section 4 gives details of our consideration on inference for TWC. Section 5 presents our qualitative and quantitative experiments. Section 6 concludes with a summary and briefly mentions future work.

2 Topic Models for Collocation Discovery

Since Chinese word discovery is similar to English collocation discovery, we first review some related topic models for collocation discovery.

Although collocation discovery has long been studied, most methods are based on frequency or variance. LDACOL is an attempt to model collocations in a topical scheme. Starting from the LDA topic model, LDACOL introduces special random variables $\vec{x}$. Variable $x_i = 1$ implies that the corresponding word w_i and previous word w_{i-1} belong to the same phrase, while $x_i = 0$ implies otherwise. Thus, LDACOL can decide the length of a phrase dynamically.

TNG is a powerful generalization of LDACOL. Its graphical model is shown in Figure 1.

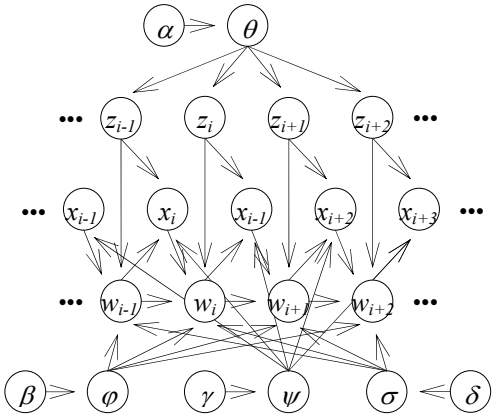

Figure 1: Topical n-gram model.

The model is defined in terms of three sets of variables: a sequence of words $\vec{w}$, a sequence of topics $\vec{z}$, and a sequence of indicators $\vec{x}$. TNG assumes the following generative process for documents.

1. For each document d, draw $\theta_d \sim$ Dirichlet(α).
2. For each topic z, draw $\varphi_z \sim$ Dirichlet(β).
3. For each topic z and each word w, draw $\sigma_{zw} \sim$ Dirichlet(δ).
4. For each topic z and each word w, draw $\psi_{zw} \sim$ Beta(γ).
5. For each word $w_{d,,i}$ in document d:
   (a) draw $x_{d,,i} \sim$ Bernoulli($\psi_{z_{d,i-1},w_{d,i-1}}$),
   (b) draw $z_{d,,i} \sim$ Discrete(θ_d),
   (c) draw $w_{d,,i} \sim$ Discrete($\varphi_{z_{d,i}}$) if $x_{d,,i}=0$,
   draw $w_{d,,i} \sim$ Discrete($\sigma_{z_{d,i},w_{d,i-1}}$) if $x_{d,,i}=1$,

where α, β, δ are Dirichlet priors and γ is a Beta prior, $z_{d,i}$ denotes the i^{th} topic assignment in document d, $w_{d,i}$ denotes the i^{th} word in document d, and $x_{d,i}$ denotes the indicator between $w_{d,i-1}$ and $w_{d,i}$. Note that the variable $x_{d,i} = 1$ implies that word $w_{d,i-1}$ and its neighbor $w_{d,i}$ belong to the same phrase, while $x_{d,i} = 0$ implies otherwise. However, the topics assigned to them ($z_{d,i-1}$ and $z_{d,i}$) are not required to be identical to each other. To decide the topic of a phrase, we can simply take the first (or last) word's topic or the most common topic in the phrase. The authors of TNG prefer to choose the last word's topic as the phrase topic because the last noun in a collocation is usually the "head noun" in English.

However, this simple strategy may be ineffective when we apply TNG to Chinese documents. The topics of "比赛" (game) and "锦标赛" (tournament) should be represented by their last characters while those of "农民" (farmer) and "农业" (agriculture) should be represented by their first characters. And occasionally, the topic of a Chinese word is not identical to any topic of its component characters. For example "蓝牙" (Bluetooth) is neither a color nor a tooth.

To overcome the limitation of TNG, we must discard its underlying assumption: that the topic of a whole word is the same as the topic of at least one of its components.

3 Modeling Word Topic and Character Topic

This section describes our topical word-character model (TWC), which models two distinct types of topics: *word topic* and *character topic*.

3.1 Word topic and character topic

To solve the problem associated with the "蓝牙" (Bluetooth) example, we need to distinguish between the topics of characters and words. Therefore, we introduce a new type of topic for words in addition to the topics assigned to characters. When generating a Chinese character, we first draw a word topic and then choose a character topic. A schematic description of this model is shown in Figure 2.

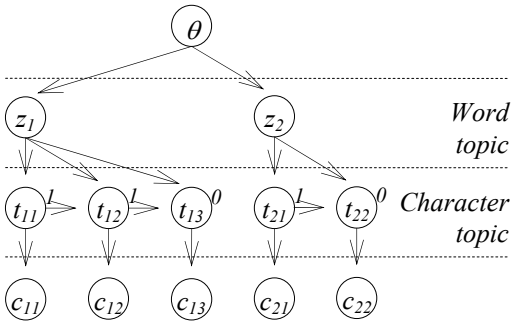

Figure 2: Schematic description of modeling Chinese documents with character and word topics.

Here, we use random variables $\vec{z}$ and $\vec{t}$ to denote word and character topics, respectively. Note that the *word topic* and *character topics* have a hierarchical tree-like structure (upper layer in Figure 2), whereas *character topics* and *characters* form a hidden Markov model (HMM) (lower layer in Figure 2).

3.2 Topical word-character model (TWC)

There are some indicators in the upper-right corner of each character topic in Figure 2. They help us to tell whether the current character belongs to the same word as the previous one. Now the question left is how to probabilistically draw these indicators, i.e., how to determine the length of the Markov chain.

There are two ways to set the values of the indicators. One is similar to that applied in the hidden semi-Markov model (HSMM), which generates the duration of a segment from the state. Accordingly, we could first choose the length of a word from the distribution associated with the word topic and then assign *0* or *1* to each indicator. The other method is to directly draw indicators from the distribution associated with the previous character and topic, just as LDACOL and TNG do. The difference between these two methods is that the former determines the length of a word in advance while the latter increases the length dynamically.

We prefer the second choice because it takes into consideration a lot of context information. In fact, our experimental results indicate that it has better performance.

The formal definition of our model with word and character topics is as follows.

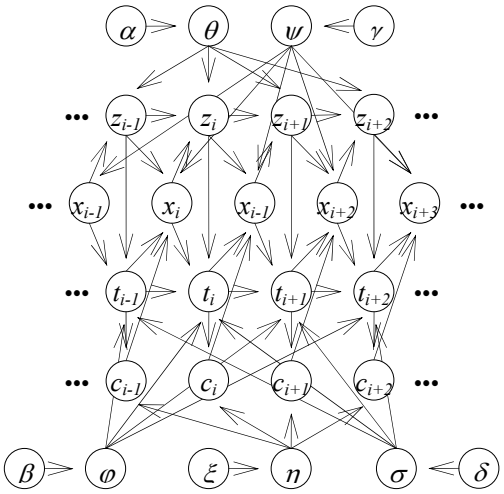

Figure 3: Topical word-character model.

TWC has four sets of variables: a sequence of characters $\vec{c}$, a sequence of character topics $\vec{t}$, a sequence of word topics $\vec{z}$, and a sequence of indicators $\vec{x}$. A document is generated via the following procedure.

1. For each document d, draw $\theta_d \sim$ Dirichlet(α);
2. For each word topic z, draw $\varphi_z \sim$ Dirichlet(β);
3. For each word topic z and each character topic t, draw $\sigma_{zt} \sim$ Dirichlet(δ);
4. For each word topic z, each character topic t and each character c, draw $\psi_{ztc} \sim$ Beta(γ);
5. For each character topic t, draw $\eta_t \sim$ Dirichlet(ξ);
6. For each character $c_{d,,i}$ in document d:
   (a) draw $x_{d,,i} \sim$ Bernoulli($\psi_{z_{d,i-1},t_{d,i-1},c_{d,i-1}}$);
   (b) draw $z_{d,,i} \sim$ Discrete (θ_d) if $x_{d,,i}=0$;
   $z_{d,,i}= z_{d,,i-1}$ if $x_{d,,i}=1$;
   (c) draw $t_{d,,i} \sim$ Discrete($\varphi_{z_{d,i}}$) if $x_{d,,i}=0$;
   draw $t_{d,,i} \sim$ Discrete($\sigma_{z_{d,i},t_{d,i-1}}$) if $x_{d,,i}=1$;
   (d) draw $c_{d,,i} \sim$ Discrete($\eta_{t_{d,i}}$).

Here, α, β, δ, η are Dirichlet priors and γ is a Beta prior, $z_{d,i}$ denotes the i^{th} word topic assignment in document d, $t_{d,i}$ denotes the i^{th} character topic assignment in document d, $c_{d,i}$ denotes the i^{th} character in document d, and $x_{d,i}$ denotes the indicator between $c_{d,i-1}$ and $c_{d,i}$.

Note that compared with the schematic model in Figure 2, each character has its corresponding

word topic in the TWC model. This is because we cannot decide how many words there will be in a document and how many characters there will be in a certain word in advance. In other words, the structure of the ideal model is not fixed. Therefore, we duplicate word topic variables for each character.

4 Inference with TWC

Many approximate inference techniques such as variational methods, expectation propagation, and Gibbs sampling can be applied to graphical models. We use Gibbs sampling to perform our Bayesian inference in TWC.

Gibbs sampling is a simple and widely applicable Markov chain Monte Carlo (MCMC) algorithm. In a traditional procedure, variables are sequentially sampled from their distributions conditioned on all other variables in the model.

An extension of the basic approach is to choose blocks of variables first and then sample jointly from the variables in each block in turn, conditioned on the remaining variables; this is called blocking Gibbs sampling.

When sampling for TWC, we separate variables into three types of blocks in the following manner (as shown in Figure 4).

1. character variables t_i

2. indicators x_i, whose value is *1* after *n* iterations

3. word topics z_i, z_{i+1}, …, z_{i+l-1} and indicator x_i, satisfying $x_i=x_{i+l}=1$ and $x_j=0$ (j from i to $i+l-1$) after *n* iterations

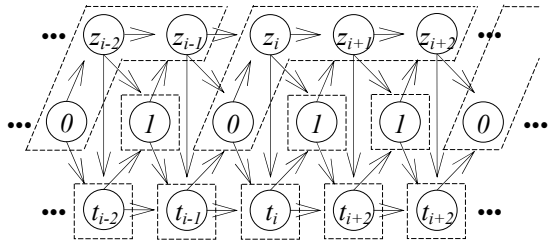

Figure 4: Illustration of partition in a certain iteration.

Note that variables $\vec{\theta},\vec{\varphi},\vec{\sigma},\vec{\psi}$ and $\vec{\eta}$ are not sampled. This is because we can integrate them out according to their conjugate priors. We only need to sample variables $\vec{z}$, $\vec{x}$, and $\vec{t}$.

Before discussing the inference of conditional probabilities, let us analyze our partition strategy in detail. We will explain the reasons for (1) sampling z_i, z_{i+1}, …, z_{i+l-1} together and (2) sampling $\vec{z}$ and x_i together

1. Why do we sample z_i, z_{i+1}, …, z_{i+l-1} together?

Assume that we draw z_i, z_{i+1}, …, z_{i+l-1} one by one, and it is now time to sample z_{i+1} according to the conditional probability

$$P(z_{i+1}=j \mid \vec{z}_{-(i+1)},\vec{x},\vec{t},\vec{c},\alpha,\beta,\delta,\gamma,\xi),$$

where $\vec{z}_{-(i+1)}$ denotes a word topic except z_{i+1}.

Recall step 6-b in the generative TWC model: it says "If $x_{d,,i}=1$, then $z_{d,,i}= z_{d,,i-1}$", which implies

$$P(z_{i+1} \mid z_i, x_{i+1}=1)=I(z_{i+1}=z_i).$$

As this probability is a factorial of the target probability, it follows that

$$P(z_{i+1}=j \mid \vec{z}_{-(i+1)},\vec{x},\vec{t},\vec{c},\alpha,\beta,\delta,\gamma,\xi)=0$$

for all $j \neq z_i$. In other words, z_{i+1} should be equal to z_i and not change during sampling.

It seems that step 6-b in the generative model causes the problem. But supposing that we do not set z_{i+1} to z_i; it is still more reasonable to sample $\vec{z}$ together. According to our partition principle, x_i, x_{i+1}, …, x_{i+l-1}, x_{i+l} is a continuous indicator sequence whose head and tail are both *0* and the rest are *1*, which implies that character string c_i, c_{i+1}, …, c_{i+l-1} forms a word and has the same word topic. Recall the schematic model in Figure 2: the *word topic* and *character topics* have a tree-like structure and each word has only one word topic node. We add some auxiliary duplicates just because the ideal model is not fixed. Therefore, it is natural to sample the word topic together with its duplicates.

2. Why do we sample $\vec{z}$ and x_i together?

Let us consider the probability of converting x_i from *0* to *1* in the current sampling iteration. Assume that the number of word topics is *3*, $z_{i-1}=2$, and

$$P(z_i=\ldots=z_{i+l-1}=j \mid x_i=0)=1/3 \quad (1\le j\le 3),$$

$$P(x_i=k \mid z_i=\ldots=z_{i+l-1}=2)=1/2 \quad (0\le k\le 1),$$

where other variables and priors are omitted. If we first sample $\vec{z}$ and next sample x_i, then the probability of drawing *1* for x_i is *1/6*, according to the multiplication principle. If we sample $\vec{z}$ and x_i together, the probability of drawing *1* for x_i is $P(z_i=\ldots=z_{i+l-1}=2,x_i=1)$.

Since

$$1=\sum_{k=0}^{1}\sum_{j=1}^{3}P(z_i=\ldots=z_{i+l-1}=j,x_i=k)$$

$$=\sum_{j=1}^{3}P(z_i=\ldots=z_{i+l-1}=j,x_i=0)$$

$$+(z_i=\ldots=z_{i+l-1}=2,x_i=1) \quad (z_{i-1}=2)$$

and

$$P(z_i=\ldots=z_{i+l-1}=2,x_i=1)$$

$$= P(z_i = ... = z_{i+l-1} = 2, x_i = 0)$$
$$= P(z_i = ... = z_{i+l-1} = j, x_i = 0) \quad (1 \le j \le 3)'$$

we get

$$P(z_i = ... = z_{i+l-1} = 2, x_i = 1) = 1/4 > 1/6 .$$

In conclusion, the model is more likely to form long words, if we sample $\vec{z}$ and x_i together. This is preferred because both TNG and TWC tend to produce shorter words than we would like.

For each type of block, we need to work out the corresponding conditional probability.

$$P(t_{d,i} = s \mid \vec{z}, \vec{x}, \vec{t}_{d,-i}, \vec{c}, \alpha, \beta, \delta, \gamma, \xi)$$
$$P(x_{d,i} = k \mid \vec{z}, \vec{x}_{d,-i}, \vec{t}, \vec{c}, \alpha, \beta, \delta, \gamma, \xi)$$
$$P(z_{d,i} = z_{d,i+1} = \cdots = z_{d,i+l-1} = j, x_{d,i} = k \mid \vec{z}_{d,-(i:i+l-1)}, \vec{x}_{d,-i}, \vec{t}, \vec{c}, \alpha, \beta, \delta, \gamma, \xi)$$

where $\vec{t}_{d,-i}$ denotes the character topic assignments except $t_{d,i}$, $\vec{x}_{d,-i}$ denotes the indicators except $x_{d,i}$, and $\vec{z}_{d,-(i:i+l-1)}$ denotes the word topic assignments except $\vec{z}_{d,j}$ (j from i to $i+l-1$). Details of the derivation of these conditional probabilities are provided in Appendix A.1.

5 Experiments

In this section, we discuss our evaluation of TWC in Chinese document modeling and Chinese word and topic discovery.

5.1 Modeling documents

To evaluate the generalization performance of our model, we trained both TWC and TNG on a Chinese corpus and computed the perplexity of the held-out test set. Perplexity, which indicates the uncertainty in predicting a single character, is a standard measure of performance for statistical models of natural language. A lower perplexity score indicates better generalization performance.

Formally, the perplexities for TWC and TNG are defined as follows.

$$perplexity_{TWC}(D_{test}) = \exp\{-\frac{\sum_{d=1}^{D} \log p(\vec{c}_d \mid \hat{\vec{\theta}}, \hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}}, \hat{\vec{\eta}})}{\sum_{d=1}^{D} N_d}\},$$

where D_{test} is the testing data, D is the number of documents in D_{test}, N_d is the number of characters in document d, $\hat{\theta}_{d,z}$ is simply set to $1/Z$ (Z is number of word topics), and $\hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}}, \hat{\vec{\eta}}$ are posterior estimates provided by applying TWC to training data. Details of the parameter estimation for TWC are provided in Appendix A.2.

$$perplexity_{TNG}(D_{test}) = \exp\{-\frac{\sum_{d=1}^{D} \log p(\vec{c}_d \mid \hat{\vec{\theta}}, \hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}})}{\sum_{d=1}^{D} N_d}\},$$

where D_{test}, D, and N_d are the same as defined for the TWC perplexity, $\hat{\theta}_{d,z}$ is simply set to $1/T$ (T is number of topics), and $\hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}}$ are posterior estimates provided by applying TNG to training data.

Now, the remaining question is how to work out the likelihood function in the definition of perplexity. The likelihood function can be obtained by marginalizing latent variables, but the time complexity is exponential. Therefore, we propose an efficient method of computing the likelihood that is similar to the *forward* algorithm for an HMM. Details of the *forward* approach to computing likelihood for TWC and TNG are provided in Appendix B.

In our experiments, we used a subset of Chinese corpus LDC2005T14. The dataset contains 6000 documents with 4476 unique characters and 2,454,616 characters. We evaluated both TWC and TNG using 10-fold cross validation. In each experiment, both models ran for 500 iterations on 90% of the data and computed the complexity for the remaining 10% of the data.

TWC used fixed Dirichlet (Beta) priors $\alpha=1$, $\beta=1$, $\delta=1$, $\gamma=0.1$ and $\xi=0.01$ while TNG used $\alpha=1$, $\beta=0.01$, $\delta=0.01$, and $\gamma=0.1$.

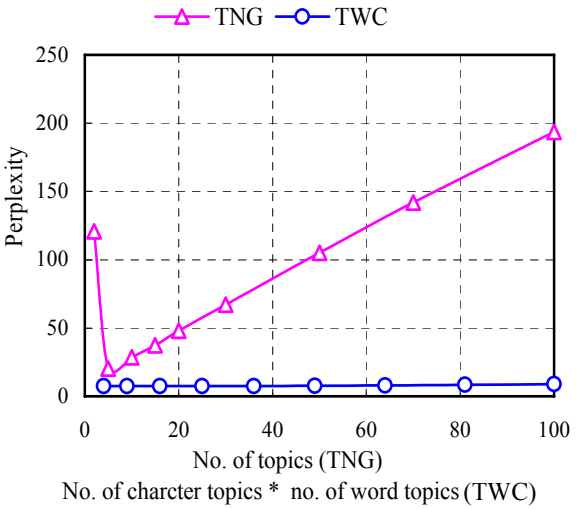

Figure 5: Perplexity results with LDC2005T14 corpora for TNG and TWC.

The results of these computations are shown in Figure 5. Note that the abscissa for TWC is the number of word topics (Z) multiplied by the

number of character topics (T), while the abscissa for TNG is the number of topics (T). They both represent the number of partitions into which the model classifies characters.

Chance performance results in a perplexity equal to the number of unique characters, which was 4476 in our experiments. Therefore, both TWC and TNG are competitive models. And the lower curve shows that TWC is much better than TNG.

We also found that both perplexity curves increased with the number of partitions. In other words, both models suffer from overfitting issues. This is because the documents in a test set are very likely to contain words that do not appear in any of the documents in the training set. Such words will have a very low probability, which is inversely proportional to the number of partitions. Therefore, the perplexity of TWC increased from 7.3513 ($Z*T=2*2$) to 8.9953 ($Z*T=10*10$), while that of TNG increased from 20.3789 (T=5) to 193.6065 ($T=100$).

5.2 Chinese word and topic discovery

As shown in the previous subsection, TWC is a competitive method for topically modeling Chinese documents. Next, we show its ability to extract Chinese words and topics in comparison with TNG.

In our qualitative experiments, the task of Chinese word and topic discovery was addressed as a supervised learning problem, where a set of words with their topical assignments was given as a seed set. Each seed can be viewed as some constraints imposed on the TWC and TNG models. For example, suppose that "教师" (teacher) together with its assignment "education" is a seed. This assumption implies that the indicator between characters "教" and "师" is *1* and that the (word) topic for each character is "education".

We make use of such constraints in a simple but effective way. In each sampling iteration, we first sample all variables as usual and then reset observed variables according to the constraints.

We used 8000 Chinese documents in the Chinese Gigaword corpus (LDC2005T14) provided by the Linguistic Data Consortium for our experiments. The dataset contains 4651 unique characters and 3,295,810 characters.

The number of word topics in TWC, the number of character topics in TWC, and the number of topics in TNG were all set to 15. Furthermore, 16 seeds scattered in 4 distinct topics were given, as listed in Table 1 column "seed". Dirichlet (Beta) priors were set to the same values as described in the previous subsection.

Word and topic assignments were extracted after running 300 Gibbs sampling iterations on the training corpus together with the seed set. For the TNG model, we took the first character's topic as the topic of the word. We omitted one-character words and ranked the extracted words using the following formula

$$\frac{occ_i(W)}{\sum_{i=1}^{15} occ_i(W)},$$

where $occ_i(W)$ represents how many words were assigned to (word) topic i. The top-50 extracted words are presented in Table 1.

We find found that both TWC and TNG could assemble many common words used in corresponding topics. And the TWC model had advantages over the TNG model in the following three respects.

First, TNG drew more words related to the seeds. In Table 1, highly related words are marked in pink (underline) and partly related words are marked in blue (italic). It is clear that the TWC column is more colorful than the TNG column.

Secondly, we found that many words extracted by TNG had the same prefix. For example, consider the topic "agriculture": there are 14 words marked with superscript 1 in Table 1. They all have the prefix "农". This is because we took the first character's topic as the topic of the word. Although this strategy is beneficial in some cases, such as for words with prefix "农", it is detrimental in other cases. For example, "甘蔗" (sugar cane) and "甘肃" (Gansu) have the same prefix and topic assignment, but the latter is a name of a province in China and is not related to agriculture. Similarly, even though the character string "伊外" does not form a Chinese word, this string "伊外" and "伊朗" (Iran) are classified in the same cluster. Compared with TNG, TWC can also extract words whose topics are identical to the topic of any character. For example, the topics "自由泳" (freestyle swimming), "混合泳" (medley swimming), and "蝶泳" (butterfly stroke) depend on their suffixes.

Thirdly, although TNG stands for "topical n-gram model", it infrequently draws words containing more than two characters. On the other hand, the TWC model extracts many n-character words, such as "美国总统布什" (president of United States, George Bush), "个人混合泳" (individual medley), and "百分之四" (four per-

Seeds	TNG	TWC
足球(football) 球员(player) 比赛(match) 冠军 (championship)	撑竿, 乒乓, 比赛, 球员, 足球, 冠军, 选手, 摔跤, *世乒*, 淘汰, 训练, 纪录, 选拔, *世界*, 宇宙, 选择, 分钟, *朝鲜*, 今晚, 威廉, 分别, 邀请, 分之, 世纪, 分裂, 辽宁, 分歧, 姑娘, 纪念, 香港, 今天, 结果, 公斤, 结束, 今年, 北京, 公里, 分析, 威胁, 毛腿, 分配, *开幕*, 沈阳, 分子, 开始, 记者, 负责, *蒙古*, 今后, *保持*	蝶泳, *标赛* [2], 上届, 自由泳, 混合泳, 比分, 总分, 冠军, 接力, 单打, 蛙泳, 比赛, 赵剑, 大师, 女子, 第一名, 速滑, 本届, 选手, 苏联队, 冰球, 亚军, 游泳, 国队 [2], 决赛, 第三名, 两枚, 第七届, 国选手 [2], 个人混合泳, 全能, 金牌, 男子, 战胜, 谢军, 夺得, 银牌, *两项*, 名列, 公斤级, 冬运会, 训练, *苏联*, 美分, 领先, *获得*, 成绩, 目的, 运动员, 第三
粮食 (foodstuff) 农村 (country) 农民 (farmer) 水稻 (paddies)	粮食, 农贸 [1], 水稻, 农林 [1], 橄榄, 农副 [1], 农药 [1], 农户 [1], 蘑菇, 农膜 [1], 农村 [1], 灌溉, 农牧 [1], 农闲 [1], 蔬菜, 潍坊, 农业 [1], 农机, 农奴 [1], 玫瑰, 农民 [1], 农垦 [1], 柑桔, 农田 [1], 挂钩, 鞠躬, 甘蔗, 甘薯, 覆盖, 甘肃, 笼罩, 帐篷, 土壤, 偏僻, 灿烂, 左右, 拉玛, 拉萨, *公斤*, 帮助, 丝绸, 公路, 群众, *公顷*, 公里, 百姓, 沙漠, 蒙古, 积累, *培育*	农村, *推广*, 农民, 粮食, 水稻, 首蓿, 秸秆, 流通, 农户, 农产, 乡镇, *社会化*, 土地, 丰收, 灌溉, 妇女, 农业, 增产, 责任制, *基层*, *万只* [2], *全省*, *亿公斤* [2], *试验*, *百分之八*, *多万* [2], 反映, *科技*, 各地, 当地, *多公斤* [2], *集体*, 服务, *承包*, *公斤*, *负担*, *配套*, *万公斤* [2], 生产, 投入, *全县*, *亿亩* [2], *万户* [2], 生活, *百分之四*, 生育, *销售*, 总产, *人均*, 年产
学校 (school) 教师 (teacher) 学生 (student) 教育 (education)	叮嘱, 学生, 教训, 教徒, 遐迩, 孤寡, 学校, 胳膊, 腼腆, 教界, 喇叭, 教材, 学雷锋, 荟萃, 教授, 学儒, 慷慨, 教师, 教练, 蜘蛛, *钥匙*, 凛冽, 骄傲, 教育, 歹徒, 巾帼, 瞻仰, 残疾, 雷锋, 学谦, 雷洁, 裕禄, 烹饪, 姑娘, 残废, 舞蹈, 兄弟, 旗帜, 儿童, 精彩, 学习, 遗址, 兢兢, 呼唤, 纪念, 李瑞, 群岛, 记载, 精锐, 精神	学校, 学生, 教育, 教师, *国防*, *民族*, 宣传, 毕业, 学习, 经费, 培养, 职工, 纪律, 知识, 家庭, 重视, 实施, 教材, 事业, 培训, 社会主义, 文化, 进一步, 工作, 只有, 保证, 继续, *活动*, 任务, 推动, 小标题, 帮助, 通过, 通讯, *思想*, 加强, *研究*, 记者, 考试, 开始, 要求, 先后, 期间, *开展*, 方面, 问题, 精神, *技术*, 地区, 成绩
战争 (war) 士兵 (solider) 将军 (general) 武器 (weapon)	伊外, 伊朗, *朝圣*, 伊斯, 朝柱, 巴勒, 士兵, 拉维, 巴基, 威尔, 拉曼, 摧毁, 伊始, 伊利, 朝觐, 巴嫩, 摧残, 伊奇, 轰炸, 巴塞, 伊格, 海盗, 海夫, 巴格, 巴黎, 巴乔, 轰鸣, 巴西, 轰击, 武器, 伊沙, 拉彻, 避难, 拉法, 葡萄, 将军, 战壕, 伊拉, 威者, 战舰, 战双, 拉脱, 伊军, 战飞, 拉姆, 伊拉克, 拉瓜, 战争, 拉底, 伊拉姆	战争, 将军, 武器, 士兵, *钱其琛*, 轰炸, 导弹, 撤军, 美军, *美国总统布什*, 伊军, *讨论海湾*, *今天进入第* [2], *对海湾* [2], 科威特, *综合报道*, 海湾, 危机, 多国部队, 爆发, *新闻分析*, 死亡, *地面*, 结束后, *之一*, 摧毁, 综述, *结束*, *爆发以来*, 战舰, 伊拉克, *号决议* [2], 表示, *万桶*, 提供, 造成, *宣布*, *美国*, 包括, 问题, 目前, 同时, 今年, 公里, 去年, 其中, *美元*, 亿元, 万元, 万吨

Table 1: Top-50 words extracted by TWC and TNG.

cent). This is partly due to our sampling strategy, discussed in subsection 4.1, which increases the probability of forming long words.

We also found that some extracted character strings were very close to real Chinese words. For instance, "标赛" is a substring of "锦标赛" (tournament); "国选手" is a suffix of "中国选手" (Chinese player), "美国选手" (American player), and "法国选手" (French player); and "万公斤" is a substring of "十万公斤" (100 thousand kilograms) and "五万公斤" (50 thousand kilograms). (Such substrings are marked with superscript 2 in Table 1.) We believe that this result occurred because the training corpus was not large enough and that TWC will achieve better performance with a large dataset.

6 Conclusion and Future Work

In this paper, we presented a topical word-character (TWC) model, which models two distinct types of topics: *word topic* and *character topic*. The experimental results show that TWC is a powerful approach to modeling Chinese documents according to the standard evaluation measure of *perplexity*. We also demonstrated TWC's ability to detect words and assign topics.

Since TWC is a straightforward improvement that removes the limitations of existing topical collocation models, we expect that its application to English collocation will also result in higher performance.

Appendix A.1 Gibbs Sampling Derivation for TWC

Symbols used here are defined as follows.

C is the number of unique characters, T is the number of character topics, and Z is the number of word topics.

N_d denotes the number of characters in a document d.

$I(\cdot)$ is an indicator function, taking the value 1 when its argument is true, and 0 otherwise.

$q_{d,z,0}$ represents how words are assigned to topic z in document d; $p_{z,t,c,k}$ represents how many times an indicator is k given the previous character c, the previous character topic t, and the previous word topic z; $n_{z,t,0}$ represents how many times a character topic is t given a word topic z and the corresponding indicator 0; $m_{z,v,t,1}$ represents how many times a character topic is t given a word topic z, the previous character topic v, and the corresponding indicator 1; and $r_{t,c}$ represents how many times character c is assigned to character topic t.

$$P(t_{d,i} = s \mid \vec{z}, \vec{x}, \vec{t}_{d,-i}, \vec{c}, \alpha, \beta, \delta, \gamma, \xi)$$

$$\propto \int \prod_{d'=1}^{D} P(\vec{\theta}_{d'} \mid \alpha) \cdot \prod_{d'=1}^{D} \prod_{i'=1}^{N_{d'}} P(z_{d',i'} \mid x_{d',i'}, z_{d',i'-1}, \vec{\theta}_{d'}) \cdot d\vec{\theta} \times$$

$$\int \prod_{z'=1}^{Z} \prod_{t'=1}^{T} \prod_{c'=1}^{C} P(\vec{\psi}_{z',t',c'} \mid \gamma) \cdot \prod_{d'=1}^{D} \prod_{i'=1}^{N_{d'}} P(x_{d',i'} \mid \vec{\psi}_{z_{d',i'-1}, t_{d',i'-1}, c_{d',i'-1}}) \cdot d\vec{\psi}$$

$$\times \iint \prod_{z'=1}^{Z} P(\vec{\varphi}_{z'} \mid \beta) \cdot \prod_{z'=1}^{Z} \prod_{t'=1}^{T} P(\vec{\sigma}_{z',t'} \mid \delta) \cdot \prod_{d'=1}^{D} \prod_{i'=1}^{N_{d'}} P(t_{d',i'} \mid x_{d',i'}, \vec{\varphi}_{z_{d',i'}},$$

$$\vec{\sigma}_{z_{d',j'},t_{d',j'-1}}) \cdot d\vec{\varphi} \cdot d\vec{\sigma} \times \int \prod_{t'=1}^{T} P(\vec{\eta}_{t'} \mid \xi) \cdot \prod_{d'=1}^{D} \prod_{i'=1}^{N_{d'}} P(c_{d',i'} \mid \vec{\eta}_{t_{d',i'}}) \cdot d\vec{\eta}$$

$$\propto \int \prod_{z'=1}^{Z} \prod_{t'=1}^{T} \prod_{c'=1}^{C} \frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2} \cdot \prod_{x'=1}^{2} (\psi^{x'}_{z',t',c'})^{\gamma-1} \cdot \prod_{x'=1}^{2} (\psi^{x'}_{z',t',c'})^{p_{z',t',c',x'}} \cdot$$

$$\psi^{x_{d,j+1}}_{z_{d,j},s,c_{d,j}} \cdot d\vec{\psi} \times \int \prod_{t'=1}^{T} \frac{\Gamma(C\xi)}{\Gamma(\xi)^C} \cdot \prod_{c'=1}^{C} (\eta^{c'}_{t'})^{\xi-1} \prod_{c'=1}^{C} (\eta^{c'}_{t'})^{r_{t',c'}} \cdot \eta^{c_{d,j}}_{s} d\vec{\eta} \times$$

$$\iint \prod_{z'=1}^{Z} \frac{\Gamma(T\beta)}{\Gamma(\beta)^T} \cdot \prod_{t'=1}^{T} (\varphi^{t'}_{z'})^{\beta-1} \cdot \prod_{t'=1}^{T} (\varphi^{t'}_{z'})^{n_{z',t',0}} \cdot \prod_{z'=1}^{Z} \prod_{t'=1}^{T} \frac{\Gamma(T\delta)}{\Gamma(\delta)^T} \cdot$$

$$\prod_{v'=1}^{T} (\sigma^{v'}_{z',t'})^{\delta-1} \cdot \prod_{v'=1}^{T} (\sigma^{v'}_{z',t'})^{m_{z',t',v',1}} \cdot \begin{cases} \varphi^{s}_{z_{d,j}} & x_{d,i}=0 \\ \sigma^{s}_{z_{d,j},t_{d,j-1}} & x_{d,i}=1 \end{cases} \cdot d\vec{\varphi} \cdot d\vec{\sigma}$$

$$\propto \frac{p_{z_{d,j},s,c_{d,j},x_{d,j}} + \gamma}{p_{z_{d,j},s,c_{d,j},*} + 2\gamma} \times \frac{r_{s,c_{d,j}} + \xi}{r_{s,*} + C\xi} \times \begin{cases} \dfrac{n_{z_{d,j},s,0} + \beta}{n_{z_{d,j},*,0} + T\beta} & x_{d,i}=0 \\ \dfrac{m_{z_{d,j},t_{d,j-1},s,1} + \delta}{m_{z_{d,j},t_{d,j-1},*,1} + T\delta} & x_{d,i}=1 \end{cases}$$

Similarly,

$$P(x_{d,i}=k \mid \vec{z}, \vec{x}_{d,-i}, \vec{t}, \vec{c}, \alpha, \beta, \delta, \gamma, \xi)$$

$$\propto \begin{cases} \dfrac{q_{d,z_{d,j},0} + \alpha}{q_{d,*,0} + Z\alpha} & k=0 \\ I(z_{d,i}=z_{d,i-1}) & k=1 \end{cases} \times \frac{p_{z_{d,j-1},t_{d,j-1},c_{d,j-1},k} + \gamma}{p_{z_{d,j-1},t_{d,j-1},c_{d,j-1},*} + 2\gamma} \times$$

$$\begin{cases} \dfrac{n_{z_{d,j},t_{d,j},0} + \beta}{n_{z_{d,j},*,0} + T\beta} & k=0 \\ \dfrac{m_{z_{d,j},t_{d,j-1},t_{d,j},1} + \delta}{m_{z_{d,j},t_{d,j-1},*,1} + T\delta} & k=1 \end{cases}$$

$$P(z_{d,i}=z_{d,i+1}=\cdots=z_{d,i+l-1}=j, x_{d,i}=k \mid \vec{z}_{d,-(i:i+l-1)}, \vec{x}_{d,-i}, \vec{t}, \vec{c}, \alpha, \beta, \delta, \gamma, \xi)$$

$$\propto \begin{cases} \dfrac{q_{d,j,0} + \alpha}{q_{d,*,0} + Z\alpha} & k=0 \\ I(z_{d,i-1}=j) & k=1 \end{cases} \times \frac{p_{z_{d,j-1},t_{d,j-1},c_{d,j-1},k} + \gamma}{p_{z_{d,j-1},t_{d,j-1},c_{d,j-1},*} + 2\gamma} \cdot$$

$$\prod_{u=2}^{l+1} \frac{p_{j,t_{d,j+u-2},c_{d,j+u-2},x_{d,j+u-1}} + \gamma}{p_{j,t_{d,j+u-2},c_{d,j+u-2},*} + 2\gamma} \times \prod_{u=2}^{l} \left(\frac{m_{j,t_{d,j+u-2},t_{d,j+u-1},1} + \delta}{m_{j,t_{d,j+u-2},*1} + T\delta} \right) \cdot$$

$$\begin{cases} \dfrac{n_{j,t_{d,j},0} + \beta}{n_{j,*,0} + T\beta} & k=0 \\ \dfrac{m_{j,t_{d,j-1},t_{d,j},1} + \delta}{m_{j,t_{d,j-1},*,1} + T\delta} & k=1 \end{cases}$$

Appendix A.2 Parameter estimation for TWC

After each Gibbs sampling iteration, we obtain posterior estimates $\hat{\theta}, \hat{\varphi}, \hat{\sigma}, \hat{\psi}$ and r by

$$\hat{\theta}_{d,z} = \frac{q_{d,z,0} + \alpha}{q_{d,*,0} + Z\alpha} \qquad \hat{\psi}_{z,t,c,k} = \frac{p_{z,t,c,k} + \gamma}{p_{z,t,c,*} + 2\gamma}$$

$$\hat{\sigma}_{z,v,t} = \frac{m_{z,v,t,1} + \delta}{m_{z,v,*,1} + T\delta} \qquad \hat{\varphi}_{z,t} = \frac{n_{z,t,0} + \beta}{n_{z,*,0} + T\beta} \quad ,$$

$$\hat{\eta}_{t,c} = \frac{r_{t,c} + \xi}{r_{t,*} + C\xi}$$

where the symbols are the same as those defined in Appendix A.1. These values correspond to the predictive distribution over new word topics, new indicators, new character topics, and new characters.

Appendix B. Likelihood Function Derivation for TWC and TNG

To compute the likelihood function for TWC, a quaternion function g_i is defined as follows: (formula has a broken character)

$$g_i(r,s,u,v) \triangleq P(c_{d,1}, c_{d,2}, \ldots, c_{d,i}, z_{d,i}=r, x_{d,i}=s, t_{d,i-1}=u, t_{d,i}=v \mid \hat{\vec{\theta}}, \hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}}, \hat{\vec{\eta}}) \quad .$$

Then, it is clear that

$$P(\vec{c}_d \mid \hat{\vec{\theta}}, \hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}}, \hat{\vec{\eta}}) = \sum_{r=1}^{Z} \sum_{s=0}^{1} \sum_{u=1}^{T} \sum_{v=1}^{T} g_{N_d}(r,s,u,v) \, ,$$

where Z is the number of word topics and T is the number of character topics. The function g_i can be rewritten in a recursive manner.

$$g_1(r,1,u,v) = 0$$

$$g_1(r,0,u,v) = \hat{\theta}^{r}_{d} \times \frac{1}{T} \times \varphi^{v}_{r} \times \eta^{c_{d,1}}_{v}$$

$$g_{i+1}(r,s,u,v) = \sum_{j=1}^{Z} \sum_{k=0}^{1} \sum_{l=1}^{T} g_i(j,k,l,u) \times \hat{\psi}^{s}_{j,u,c_{d,i}} \times \hat{\eta}^{c_{d,i+1}}_{v} \times \begin{cases} \hat{\theta}^{r}_{d} & s=0 \\ I(r=j) & s=1 \end{cases} \times \begin{cases} \hat{\varphi}^{v}_{r} & s=0 \\ \hat{\sigma}^{v}_{r,u} & s=1 \end{cases}$$

Similarly we can define function h_i to help compute the likelihood for TNG. (formula has a broken character)

$$h_i(r,s) \triangleq P(c_{d,1}, \ldots, c_{d,i}, z_i=r, x_i=s \mid \hat{\vec{\theta}}, \hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}})$$

$$P(\vec{c}_d \mid \hat{\vec{\theta}}, \hat{\vec{\varphi}}, \hat{\vec{\sigma}}, \hat{\vec{\psi}}) = \sum_{r=1}^{Z} \sum_{s=0}^{1} h_{N_d}(r,s)$$

$$h_1(r,1) = 0$$

$$h_1(r,0) = \hat{\theta}^{r}_{d} \times \hat{\varphi}^{c_{d,1}}_{r}$$

$$h_{i+1}(r,s) = \sum_{j=1}^{Z} \sum_{k=0}^{1} h_i(j,k) \times \hat{\psi}^{s}_{j,c_{d,i}} \times \hat{\theta}^{r}_{d} \times \begin{cases} \hat{\varphi}^{c_{d,i+1}}_{r} & s=0 \\ \hat{\sigma}^{c_{d,i+1}}_{r,c_{d,i}} & s=1 \end{cases}$$

References

Blei, D. M., Ng, A. Y., and Jordan, M. J. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Griffiths, T. L. and Steyvers, M. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences, 101 (suppl. 1)*, 5228-5235.

Steyvers, M. and Griffiths, T. L. 2007. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. T. 2007. Topics in Semantic Representation *Psychological Review*, 114(2), 211-244.

Wang, X., McCallum, A., and Wei, X. 2007. Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. *Proceedings of the 7th IEEE International Conference on Data Mining* (ICDM 2007).