WINGS: Writing with Intelligent Guidance and Suggestions

Xianjun Dai, Yuanchao Liu*, Xiaolong Wang, Bingquan Liu
School of Computer Science and Technology
Harbin Institute of Technology, China
{xjdai, lyc, wangxl, liubq}@insun.hit.edu.cn

Abstract

Without inspirations, writing may be a frustrating task for most people. In this study, we designed and implemented *WINGS*, a Chinese input method extended on IBus-Pinyin with intelligent writing assistance. In addition to supporting common Chinese input, *WINGS* mainly attempts to spark users' inspirations by recommending both word level and sentence level writing suggestions. The main strategies used by *WINGS*, including providing syntactically and semantically related words based on word vector representation and recommending contextually related sentences based on LDA, are discussed and described. Experimental results suggest that *WINGS* can facilitate Chinese writing in an effective and creative manner.

1 Introduction

Writing articles may be a challenging task, as we usually have trouble in finding the suitable words or suffer from lack of ideas. Thus it may be very helpful if some writing reference information, e.g., words or sentences, can be recommended while we are composing an article.

On the one hand, for non-english users, e.g., Chinese, the Chinese input method is our first tool for interacting with a computer. Nowadays, the most popular Chinese input methods are Pinyin-based ones, such as Sougou Pinyin[1] and Google Pinyin[2]. These systems only present accurate results of Pinyin-to-Character conversion. Considering these systems' lack of suggestions for related words, they hardly provide writers with substantial help in writing. On the other hand, try to meet the need of writing assistance, more and more systems facilitating Chinese writing have been available to the public, such as WenXin Super Writing Assistant[3] and BigWriter[4], and among others. However, due to their shortcomings of building examples library manually and lack of corpus mining techniques, most of the time the suggestions made by these systems are not creative or contextual.

Thus, in this paper, we present Writing with INtelligent Guidance and Suggestions (*WINGS*)[5], a Chinese input method extended with intelligent writing assistance. Through *WINGS*, users can receive intelligent, real-time writing suggestions, including both word level and sentence level. Different from existing Chinese writing assistants, *WINGS* mainly attempts to spark users' writing inspirations from two aspects: providing diverse related words to expand users' minds and recommending contextual sentences according to their writing intentions. Based on corpus mining with Natural Language Processing techniques, e.g., word vector representation and LDA model, *WINGS* aims to facilitate Chinese writing in an effective and creative manner.

For example, when using *WINGS* to type "xuxurusheng", a sequence of Chinese Pinyin characters for "栩栩如生" (vivid/vividly), the Pinyin-to-Character Module will generate "栩栩如生" and some other candidate Chinese words.

Then the Words Recommending Module generates word recommendations for "栩栩如生". The recommended words are obtained through calculating word similarities based on word vector representations as well as rule-based strategy (POS patterns).

In the Sentences Recommending Module, we first use "栩栩如生" to retrieve example sentences from sentences library. Then the topic similarities between the local context and the candidate sentences are evaluated for contextual

* Corresponding author
[1] http://pinyin.sogou.com
[2] http://www.google.com/intl/zh-CN/ime/pinyin
[3] http://www.xiesky.com
[4] http://www.zidongxiezuo.com/bigwriter_intro.php
[5] The DEB package for Ubuntu 64 and recorded video of our system demonstration can be accessed at this URL: http://yunpan.cn/Qp4gM3HW446Rx (password:63b3)

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 25–30, Baltimore, Maryland USA, June 23-24, 2014.

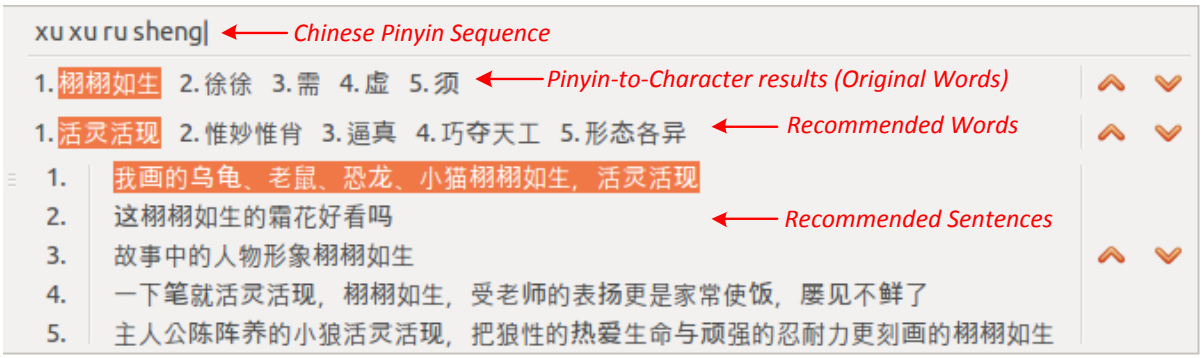

Figure 1. Screenshot of *WINGS*.

sentence recommendations.

At last in consideration of users' feedback, we introduce a User Feedback Module to our system. The recorded feedback data will in turn influence the scores of words and sentences in Recommending Modules above.

Figure 1 shows a screenshot of *WINGS*.

2 Related Work

2.1 Input Method

Chinese input method is one of the most important tools for Chinese PC users. Nowadays, Pinyin-based input method is the most popular one. The main strategy that Pinyin-based input method uses is automatically converting Pinyin to Chinese characters (Chen and Lee, 2000).

In recent years, more and more intelligent strategies have been adopted by different input methods, such as Triivi[6], an English input method that attempts to increase writing speed by suggesting words and phrases, and PRIME (Komatsu et al., 2005), an English/Japanese input system that utilizes visited documents to predict the user's next word to be input.

In our system the basic process was Pinyin → Characters (words) → Writing Suggestions (including words and sentences). We mainly focused on writing suggestions from Characters (words) in this paper. As the Pinyin-to-Character was the underlining work, we developed our system directly on the open source framework of the IBus (an intelligent input Bus for Linux and Unix OS) and IBus-Pinyin[7] input method.

2.2 Writing Assistant

As previously mentioned, several systems are available in supporting Chinese writing, such as WenXin Super Writing Assistant and Big Writer. These systems are examples of a retrieval-based writing assistant, which is primarily based on a large examples library and provides users with a search function.

In contrast, other writing assistants employ special NLP strategies. Liu et al. (2011, 2012) proposed two computer writing assistants: one for writing love letters and the other for blog writing. In these two systems, some special techniques were used, including text generation, synonym substitution, and concept expansion. PENS (Liu et al., 2000) and FLOW (Chen et al., 2012) are two writing assistants designed for students of English as a Foreign Language (EFL) practicing writing, which are mainly based on Statistical Machine Translation (SMT) strategies.

Compared with the above mentioned systems, *WINGS* is closer to retrieval-based writing assistants in terms of function. However, *WINGS* can provide more intelligent suggestions because of the introduction of NLP techniques, e.g., word vector representation and topic model.

2.3 Word Representations in Vector Space

Recently, Mikolov et al. (2013) proposed novel model architectures to compute continuous vector representations of words obtained from very large data sets. The quality of these representations was assessed through a word similarity task, and according to their report, the word vectors provided state-of-the-art performance for measuring syntactic and semantic word similarities in their test set. Their research produced the open source tool word2vec[8].

In our system, we used word2vec to train the word vectors from a corpus we processed beforehand. For the Words Recommending Module, these vectors were used to determine the similarity among different words.

[6] http://baike.baidu.com/view/4849876.htm
[7] https://code.google.com/p/ibus
[8] https://code.google.com/p/word2vec

2.4 Latent Dirichlet Allocation

The topic model Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. In this model, documents are represented as random mixtures of latent topics, where each topic is characterized by the distribution of words (Blei et al., 2003). Each document can thus be represented as a distribution of topics.

Gibbs Sampling is a popular and efficient strategy used for LDA parameter estimation and inference. This technique is used in implementing several open sourcing LDA tools, such as GibbsLDA++[9] (Phan and Nguyen, 2007), which was used in this paper.

In order to generate contextual sentence suggestions, we ensured that the sentences recommended to the user were topic related to the local context (5-10 words previously input) based on the LDA model.

3 Overview of WINGS

Figure 2 illustrates the overall architecture of *WINGS*.

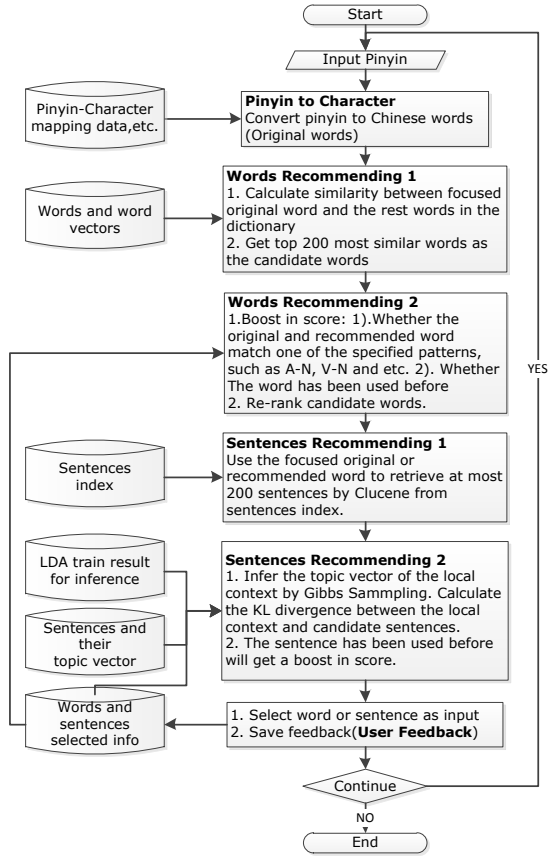

Figure 2. Overall architecture of *WINGS*.

3.1 System Architecture

Our system is composed of four different modules: **Pinyin-to-Character Module**, **Words Recommending Module**, **Sentences Recommending Module**, and **User Feedback Module**. The following sub-sections discuss these modules in detail.

3.2 Pinyin-to-Character Module

Our system is based on the open sourcing input framework IBus and extended on the IBus-Pinyin input method. Thus, the Pinyin-to-Character module is adopted from the original IBus-Pinyin system. This module converts the input Chinese Pinyin sequence into a list of candidate Chinese words, which we refer to as **original words**.

3.3 Words Recommending Module

- **Words vector representations**

In this preparatory step for word recommendation, words vector representations are obtained using the word2vec tool. This will be described in detail in Section 4.

- **Obtain the most related words**

Our system will obtain the focused **original word** and calculate the cosine similarities between this word and the rest of the words in the dictionary. Thus, we can obtain the top 200 most similar words according to their cosine values. These words are referred to as **recommended words**. According to Mikolov et al. (2013), these words are syntactically and semantically similar to the original word.

- **Re-rank the recommended words**

In order to further improve word recommending, we introduce several special POS patterns (Table 1). If the POS of the original word and the recommended word satisfy one of the POS patterns we specified, the score (based on the cosine similarity) of the recommended word will be boosted. In addition, the score of the word selected by the user before will also be boosted. Therefore, these words will be ranked higher in the recommended words list.

POS of original word	POS of recommended word
N (noun)	A (adjective)
A (adjective)	N (noun)
N (noun)	V (verb)
Any POS	Same with the original word
Any POS	L (idiom)

Table 1. Special POS patterns.

3.4 Sentences Recommending Module

- **Sentences topic distribution**

In this preparatory step for sentence

[9] http://gibbslda.sourceforge.net

recommendation, sentences topic distribution vectors and other parameters are trained using the GibbsLDA++. This step will be discussed in Section 4.

- **Retrieve relative sentences via CLucene**

The focused original or recommended word will be used to search the most related sentences in the sentences index via CLucene[10]. At most 200 sentences will be taken as candidates, which will be called **recommended sentences**.

- **Re-rank the recommended sentences**

To ensure that the recommended sentences are topic related to our local input context (5-10 words previously input), we use Gibbs Sampling to infer the topic vector of the local context, and calculate the KL divergence between the local context and each recommended sentence. Finally, the recommended sentences will be re-ranked based on their KL divergences value with respect to the local context and the boost score derived from the feedback information.

3.5 User Feedback Module

This module saves the users' feedback information, particularly the number of times when users select the recommended words and sentences. This information will be used as a boost factor for the **Words** and **Sentences Recommending Modules**. Our reasons for introducing this module are two-fold: the users' feedback reflects their preference, and at the same time, this information can somewhat indicate the quality of the words and sentences.

4 Data Pre-processing

In this section, the procedure of our data pre-processing is discussed in detail. Firstly, our raw corpus was crawled from DiYiFanWen[11], a Chinese writing website that includes all types of writing materials. After extracting useful composition examples from each raw html file, we merged all articles into a single file named **large corpus**. Finally, a total of 324,302 articles were merged into the large corpus (with a total size of 320 MB).

For words recommending, each of the articles in our large corpus was segmented into words by ICTCLAS [12] with POS tags. Subsequently, word2vec tool was used on the words sequence (with useless symbols filtered). Finally, the words, their respective vector representations and main POS tags were combined, and we built these data into one binary file.

For sentences recommending, the large corpus was segmented into sentences based on special punctuations. Sentences that were either too long or too short were discarded. Finally, 2,567,948 sentences were left, which we called **original sentences**. An index was created on these sentences using CLucene. Moreover, we segmented these original sentences and filtered the punctuations and stop words. Accordingly, these new sentences were named **segmented sentences**. We then ran GibbsLDA++ on the segmented sentences, and the Gibbs sampling result and topic vector of each sentence were thus obtained. Finally, we built the original sentence and their topic vectors into a binary file. The Gibbs sampling data used for inference was likewise saved into a binary file.

Table 2 lists all information on the resources of *WINGS*.

Items	Information
Articles corpus size	320 MB
Articles total count	324,302
Words total count	101,188
Sentences total count	2,567,948

Table 2. Resources information.

5 Experimental Results

This section discusses the experimental results of *WINGS*.

5.1 Words Recommending

The top 20 recommended words for the sample word "老师" (teacher) are listed in Table 3. Compared with traditional methods (using Cilin, Hownet, and so forth.), using the word vectors to determine related words will identify more diverse and meaningful related words and this quality of *WINGS* is shown in Table 4. With the diversity of recommended words, writers' minds can be expanded easily.

1-10: 同学(student), 上课(conduct class), 语文课(Chinese class), 语重心长(with sincere words and earnest wishes), 和蔼可亲(affability), 教导(guide), 讲课(lecture), 讲台(dais), 不厌其烦(patient), 全班(the whole class)
11-20: 下课(finish class), 一番话(remarks), 数学课(math class), 开小差(be absent-minded), 戒尺(ferule), 班主任(class adviser), 忐忑不安(restless), 记得(remember), 青出于蓝而胜于蓝(excel one's master), 听讲(listen to)

Table 3. Top 20 recommended words for "老师" (teacher).

[10] http://sourceforge.net/projects/clucene
[11] http://www.diyifanwen.com
[12] http://ictclas.nlpir.org

Words about	Words
Person	同学, 班主任, 全班
Quality	语重心长, 和蔼可亲, 不厌其烦
Course	语文课, 数学课
Teaching	教导, 讲课, 上课, 下课
Teaching facility	讲台, 戒尺
Student behaviour	听讲, 开小差, 忐忑不安
Special idiom	青出于蓝而胜于蓝
Others	记得, 一番话

Table 4. Diversity of recommended words for "老师" (teacher).

5.2 Sentences Recommending

By introducing the topic model LDA, the sentences recommended by *WINGS* are related to the topic of the local context. Table 5 presents the top 5 recommended sentences for the word "栩栩如生" (vivid/vividly) in two different local contexts: one refers to characters in books; the other refers to statues and sculptures. Most sentences in the first group are related to the first context, and most from the second group are related to the second context.

In order to assess the performance of *WINGS* in sentence recommendation, the following evaluation was implemented. A total of 10 Chinese words were randomly selected, and each word was given two or three different local contexts as above (contexts varied for different words). Finally, we obtained a total of 24 groups of data, each of which included an original word, a local context, and the top 10 sentences recommended by *WINGS*. To avoid the influence of personal preferences, 12 students were invited to judge whether each sentence in the 24 different groups was related to their respective local context. We believed that a sentence was related to its context only when at least 70% of the evaluators agreed. The Precision@10 measure in Information Retrieval was used, and the total average was 0.76, as shown in Table 6.

Additionally, when we checked the sentences which were judged not related to their respective local context, we found that these sentences were generally too short after stop words removal, and as a result the topic distributions inferred from Gibbs Sampling were not that reliable.

Context 1 is about characters in books: 故事(story), 人物(character), 形象(image), 作品(works)
1 这本书刻画了许多栩栩如生的人物 (The characters of this book are depicted vividly)
2 这本书人物描写栩栩如生，故事叙述有声有色 (The characters of this book are depicted vividly and the story is impressive narrative)
3 故事中的人物形象栩栩如生 (The characters of this story are depicted vividly)
4 他的作品情节惊险曲折人物栩栩如生结局出人意料 (His works are full of plot twists, vivid characters, and surprising endings)
5 书中的人物都被葛竞姐姐描写得栩栩如生 (The characters in the book are depicted vividly by Jing Zhuge)
Context 2 is about statues and sculptures: 塑像(statue), 雕塑(sculpture), 石刻(stone inscription), 寺庙(temple)
1 墙上绘满了威武的龙，栩栩如生 (The walls are painted with mighty and vivid dragons)
2 两侧的十八罗汉神态各异，栩栩如生 (On both sides there are standing 18 vivid Arhats with different manners)
3 大雄宝殿气势恢弘，殿内人物栩栩如生 (the Great Buddha Hall is grand and the statues there are vivid)
4 每尊都栩栩如生，活灵活现 (Each statue is vivid and lifelike)
5 檐角上各有七个栩栩如生的飞禽走兽像，它们各有其寓意 (On each of the eave angles there are 7 vivid statues of animals and birds with special meanings)

Table 5. Top 5 recommended sentences for "栩栩如生" (vivid/vividly) in two different local contexts.

Local Context	word 1	word 2	word 3	word 4	word 5	word 6	word 7	word 8	word 9	word 10
1	0.9	0.3	0.9	0.6	0.7	0.8	0.6	0.8	1.0	0.9
2	0.4	0.7	1.0	0.9	0.9	0.7	1.0	0.5	0.9	0.5
3	0.9	N/A	N/A	N/A	N/A	0.9	0.8	N/A	N/A	0.7
Average Precision@10 value of the 24 groups data								**0.76**		

Table 6. Precision@10 value of each word under their respective context and the total average.

5.3 Real Time Performance

In order to ensure the real time process for each recommendation, we used CLucene to index and retrieve sentences and memory cache strategy to reduce the time cost of fetching sentences' information. Table 7 shows the average and max responding time of each recommendation of randomly selected 200 different words (Our test environment is 64-bit Ubuntu 12.04 LTS OS on PC with 4GB memory and 3.10GHz Dual-Core CPU).

Item	Responding time
Average	154 ms
Max	181 ms

Table 7. The average and max responding time of 200 different words' recommending process

6 Conclusion and Future Work

In this paper, we presented *WINGS*, a Chinese input method extended with writing assistance that provides intelligent, real-time suggestions for writers. Overall, our system provides syntactically and semantically related words, as well as recommends contextually related sentences to users. As for the large corpus, on which the recommended words and sentences are based, and the corpus mining based on NLP techniques (e.g., word vector representation and topic model LDA), experimental results show that our system is both helpful and meaningful. In addition, given that the writers' feedback is recorded, *WINGS* will become increasingly effective for users while in use. Thus, we believe that *WINGS* will considerably benefit writers.

In future work, we will conduct more user experiments to understand the benefits of our system to their writing. For example, we can integrate *WINGS* into a crowdsourcing system and analyze the improvement in our users' writing. Moreover, our system may still be improved further. For example, we are interested in adding a function similar to Google Suggest, which is based on the query log of the search engine, in order to provide more valuable suggestions for users.

References

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3, pages 993-1022.

Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao and Jason S. Chang. 2012. FLOW: a first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 System Demonstrations*, pages 157-162.

Zheng Chen and Kai-Fu Lee. 2000. A new statistical approach to Chinese Pinyin input. In *Proceedings of the 38th annual meeting on association for computational linguistics*, pages 241-247.

Hiroyuki Komatsu, Satoru Takabayash and Toshiyuki Masui. 2005. Corpus-based predictive text input. In *Proceedings of the 2005 international conference on active media technology*, pages 75–80.

Chien-Liang Liu, Chia-Hoang Lee, Ssu-Han Yu and Chih-Wei Chen. 2011. Computer assisted writing system. *Expert Systems with Applications, 38(1),* pages 804-811.

Chien-Liang Liu, Chia-Hoang Lee and Bo-Yuan Ding. 2012. Intelligent computer assisted blog writing system. Expert Systems with Applications, 39(4), pages 4496-4504.

Ting Liu, Ming Zhou, Jianfeng Gao, Endong Xun and Changning Huang. 2000. PENS: A machine-aided English writing system for Chinese users. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 529-536.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA).