Chinese Preposition Selection for Grammatical Error Diagnosis

Hen-Hsen Huang, Yen-Chi Shao, and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{hhhuang, ycshao}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

Misuse of Chinese prepositions is one of common word usage errors in grammatical error diagnosis. In this paper, we adopt the Chinese Gigaword corpus and HSK corpus as L1 and L2 corpora, respectively. We explore gated recurrent neural network model (GRU), and an ensemble of GRU model and maximum entropy language model (GRU-ME) to select the best preposition from 43 candidates for each test sentence. The experimental results show the advantage of the GRU models over simple RNN and n-gram models. We further analyze the effectiveness of linguistic information such as word boundary and part-of-speech tag in this task.

1 Introduction

As learning Chinese has been popular world-wide, Chinese word spelling checking and grammatical error diagnosis for learners of Chinese language is recently advanced in the NLP community. In the NLP-TEA shared tasks (Yu et al., 2014; Lee et al., 2015a, Lee et al., 2015b), four types of grammatical errors including disorder, redundant, missing, and mis-selection are defined. These tasks focus on detecting and identifying grammatical errors, but not addressing error correction.

In English, selection of appropriate prepositions is a barrier to non-native language learners (Chodorow et al., 2007; Felica and Pulman, 2008). Many studies deal with the issue of preposition error detection and correction (Dale et al., 2012; Ng et al., 2013). The selection of Chinese prepositions is also challenging to non-native learners, especially for some common prepositions. (S1) and (S2) show a real case from the HSK corpus. The preposition 從 (cóng, "from") in (S2) is preferred to 在 (zài, "on/in/at") in (S1). In this paper, we investigate Chinese preposition selection.

(S1) **在** 公眾 利益 方面 來看 (**on** the view point of public interest)
(S2) **從** 公眾 利益 方面 來看 (**from** the view point of public interest)

A model trained on an error-annotated dataset benefits from learning the mapping between wrong instances and their corrected counterparts (Cahill et al., 2013). However, large error-annotated dataset for Chinese preposition error correction is still not available. In this paper, we utilize a large-scale corpus written by native Chinese speakers to deal with this problem. We adopt language models based on recurrent neural networks (RNNs) (Mikolov et al., 2011) to capture word dependencies in sentences. Cutting-edge methods like noise contrastive estimation (NCE) (Chen et al., 2015a) and gated recurrent unit (GRU) (Cho et al., 2014) are explored in RNNs. For the task of 43-way classification, i.e., to select the best preposition from 43 candidates, the gated recurrent neural network model (GRU) achieves an accuracy of 74.05% and an MRR of 83.08% on the Chinese Gigaword corpus (L1 corpus), and achieves an accuracy of 60.13% and an MRR of 72.54% on the HSK corpus (L2 corpus). An ensemble of gated

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 888–899, Osaka, Japan, December 11-17 2016.

recurrent neural network model and maximum entropy language model (GRU-ME) further improves the accuracy and the MRR of the GRU model on the Gigaword corpus to 76.71% and 84.89%.

The contribution of this paper is two-fold. From technological point of view, to the best of our knowledge, this paper is the first attempt to deal with Chinese preposition selection based on gated recurrent neural network model and an ensemble of GRU and ME. From linguistic point of view, it discusses the unique phenomena of Chinese prepositions with empirical evidence. The rest of this paper is organized as follows. Section 2 introduces the related work of grammatical error diagnosis in English and Chinese. We further introduce Chinese prepositions in Section 3. Section 4 shows L1 and L2 corpora used in this study. Section 5 presents our approach to Chinese preposition selection. Section 6 shows the experimental results. We further analyze the results in Section 7. Finally, Section 8 concludes this work.

2 Related Work

English preposition error detection has attracted much attention for years. Felice and Pulman (2007) proposed a voted perceptron classifier for disambiguating the uses of five common prepositions including "in", "of", "on", "to", and "with". In the work of Felice and Pulman (2008), error detection of nine common prepositions is tackled with the maximum entropy classifier. Chodorow et al. (2007) deal with the detection of preposition errors of non-native learners. In addition to error detection, some studies address the task of English preposition selection. Bergsma et al. (2009) propose a supervised language model for preposition correction. Tetreault et al. (2010) introduce parse features for this task. Cahill et al. (2013) propose a preposition error correction model trained on error-annotated data, and treated the revision logs of Wikipedia as a large error-annotated corpus. Xiang et al. (2013) propose a hybrid approach to deal with preposition selection. Zhang and Wang (2014) introduce a framework for English grammatical error correction using the maximum entropy language model for the replacement errors. Ramisa et al. (2015) address the task of preposition prediction for image descriptions with multimodal features. Related evaluations are covered in the shared tasks of HOO 2011 (Dale and Kilgarriff, 2011), HOO 2012 (Dale et al., 2012), CoNLL 2013 (Ng et al., 2013) and CoNLL 2014 (Ng et al., 2014).

In addition to Chinese spelling checking (Lee et al., 2015b), grammatical error detection in Chinese has been investigated recently. Wang (2011) shows common Chinese grammatical errors like missing components and error word orderings. Lin (2011), Yu and Chen (2012), and Cheng et al. (2014) focus on the detection and correction of word ordering errors in Chinese written by foreign students in the HSK corpus. Shiue and Chen (2016) determine if a Chinese sentence contains word usage errors.

In the NLP-TEA shared tasks (Yu et al., 2014; Lee et al., 2015a), detection of four grammatical errors are targeted. Lin and Chan (2014) train SVM classifiers with various bigram features. Zampiperi and Tan (2014) propose a frequency-based approach based on a large general corpus. Zhao et al. (2014; 2015) model the task of correction as machine translation in such a way that the wrong sentences are translated to correct ones. The preposition error detection is one of error cases in NLP-TEA shared tasks. However, the preposition correction is beyond the scope of NLP-TEA.

Different from previous works, our work focuses on the correction of Chinese prepositions. We aim at selecting suitable prepositions from a set of 43 common prepositions. To overcome the limitation of error-annotated dataset, we propose an unsupervised approach based on language models, which can be trained on a large scale L1 corpus without the need of annotation.

3 Chinese Prepositions

A preposition is a function word that is followed by a noun phrase to introduce a preposition phrase (PP). It indicates a relation between the noun phrase and other words within the sentence. For example, "in", "on", "to", "with", and "of" are most common prepositions in English. In the Penn Treebank 3 corpus (Marcus et al., 1999), 186 distinct English words are tagged with the part-of-speech (POS) tag "P" (i.e. preposition). In the 186 prepositions, some of them are abbreviations such as "altho" (although) and "w." (with).

In the same manner, we found a total of 288 distinct Chinese prepositions in the Chinese Treebank 8.0 corpus (Xue et al., 2013). The total occurrences of them are 54,239 in 71,369 sentences. In the 288 prepositions, 199 of them appear more than once, and 44 of them appear more than 100 times. The top three most frequent prepositions are 在 (zài, "on/in/at"), 對 (duì, "to/at"), and 從 (cóng, "from").

The CKIP group (1993) categorized Chinese prepositions into 66 types of senses[1]. For example, these words 直到 (zhídào), 迄 (qì), 等到 (děng dào), 比及 (bǐ jí), 及至 (jízhì), and 待到 (dài dào) share the same meaning of “until” when they are used as preposition. Note that some prepositions have other usages. The words 與 (yǔ) and 和 (hàn) can be used as preposition with the meaning of “with”, while they are also common conjunctions with the meaning of “and”. The word 對 (duì) can be used as preposition (to/at), noun (a pair of), adjective (right/correct), adverb (yes), and verb (be opposite/reply) with various senses. Another highly ambiguous word 給 (gěi) can be used as preposition (to/for/with), verb (give/let), and emphatic particle. Both 對 (duì) and 給 (gěi) are common Chinese words, they have multiple senses by their own, and they are interchangeable in some situations.

Some preposition/noun pairs usually collocate. For example, the combination of the noun 辦公室 (office) and the preposition 在 (zài, “in”) forms a common preposition phrase 在辦公室 (in the office). In a complicated sentence like (S3), there are eight words in-between 在 and 辦公室. Such a long distance dependency is challenging to language models.

(S3) 我 **在** 設於 上海 浦東 機場 西面 的 口岸 服務 **辦公室** (I am **in** the port service **office** located in the west of the Pudong airport, Shanghai)

In addition, the combination of the preposition and the noun varies according to semantics. For the noun 辦公室 (office) in (S4) and (S5), different prepositions are used. The ambiguity of preposition selection not only causes confusion to non-native learners, but also makes challenges in natural language processing.

(S4) 我 **在** 辦公室 上班 (I work **in** the office)
(S5) 我 **從** 辦公室 出發 (I depart **from** the office)

4 Datasets

We adopt the Tagged Chinese Gigaword (CGW) corpus 2.0[2] (Huang, 2009; Huang et al., 2008) as the L1 corpus. It contains 2,803,632 documents and 831,748,000 words with part-of-speech (POS) tags. By removing the sentences without prepositions, a total of 23,486,882 sentences covering 155 prepositions are collected from the Gigaword corpus. We randomly select 60% of sentences for training models due to the computation limitations. Additional 5,000 and 200,000 sentences are randomly selected as development data and test data, respectively. The development data is used for validation.

HSK dynamic composition corpus is adopted as the L2 corpus[3]. It collects articles written by students from foreign countries to study Chinese in Beijing Language and Culture University. Total 46 error categories range from character level, word level, sentence level, to discourse level are annotated and corrected in the HSK corpus. The CKIP segmentation system is used to perform Chinese word segmentation and POS tagging on the sentences in HSK because the POS tagging in the CGW corpus follows the CKIP style.

From the HSK corpus, total 745 sentences consisting of preposition errors are extracted from word-level grammatical error set. Limited to the small L2 dataset, these 745 sentences are used as test data only. The statistics of L1 and L2 datasets are listed in Table 1.

Dataset	Source	Number of Sentences
Training set (L1)	CGW	14,092,128
Development set (L1)	CGW	5,000
Test set (L1)	CGW	200,000
Test set (L2)	HSK	745

Table 1: Statistics of the L1 (CGW) and L2 (HSK) datasets used in experiments.

[1] http://rocling.iis.sinica.edu.tw/CKIP/tr/9305_2013%20revision.pdf
[2] https://catalog.ldc.upenn.edu/LDC2009T14
[3] http://202.112.195.192:8060/hsk/login.asp

Among the Chinese prepositions mentioned in Section 3, 43 prepositions, most of which are common words, appear in the HSK dataset. Note that 33 of them belong to the list of the top 50 prepositions in CGW. Furthermore, these 43 prepositions cover 88.61% of preposition uses in CGW. Figure 1 lists the 43 prepositions used in this work by the order of their occurrences.

在 (on/in/at) 對 (to/at) 從 (from) 用 (with/using) 以 (with/by) 跟 (with) 由 (by/from)
給 (to/for/with) 為 (for) 和 (to) 向 (to/toward) 與 (to) 像 (like) 對於 (for/regarding) 當 (at)
把 (owing to) 到 (to) 靠 (by) 於 (in/to/on/for/at/of) 被 (passive particle) 關於 (about/on)
隨著 (along) 比 (than) 根據 (according to/based on) 如 (as) 依 (according to/by) 就 (on)
針對 (against) 離 (from) 按照 (according to/by) 替 (for) 至於 (touching) 受 (passive particle)
至 (to/until) 等到 (until) 據 (according to) 按 (according to/by) 往 (to/toward) 經由 (via)
同 (with) 憑 (by) 趁 (at) 正當 (at)

Figure 1: Prepositions in the HSK corpus.

5 Methods

This work deals with the preposition selection error. Given a set of Chinese prepositions, *PS*, and a sentence $S = (x_1, x_2, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)$, where x_i is a preposition, we try to substitute x_i for each preposition *p* in *PS*, and choose the most probable preposition $\hat{p}$.

$$\hat{p} = \underset{p\in PS}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_{i-1}, p, x_{i+1}, \ldots, x_n)$$

Prepositions are a closed set of function words. Thus, it is feasible to substitute each of all prepositions for the location of a preposition. As listed in Section 4, *PS* contains 43 common prepositions which appear in both the CGW corpus and the HSK corpus. The probability of a sentence is estimated by a language model. The traditional n-gram language model is considered as the baseline in this work. The SRI language modeling toolkit (SRILM)[4], an implementation of n-gram language model, is adopted. The major disadvantage of n-gram is that the number of its parameters growths exponentially as the order of n-gram increases. As a result, the order of n-gram usually ranges between bigram and 5-gram. The n-gram model with high order is not impractical.

Recently, language models based on recurrent neural networks (RNNs) such as simple RNN, long short-term memory (LSTM), and gated recurrent unit (GRU) (Mikolov et al., 2011; Hochreiter and Jürgen Schmidhuber, 1997; Cho et al., 2014) have been shown to outperform traditional approaches in speech recognition and other applications. Figure 2 illustrates a basic recurrent neural network. The hidden state s_t is passed to the next step for the following update.

$$s_t = f(Wx_t + Us_{t-1})$$

where *f* is an activation function such as a sigmoid function or a tanh function, and *W* and *U* are parameters to be learned. In other words, the history information is kept.

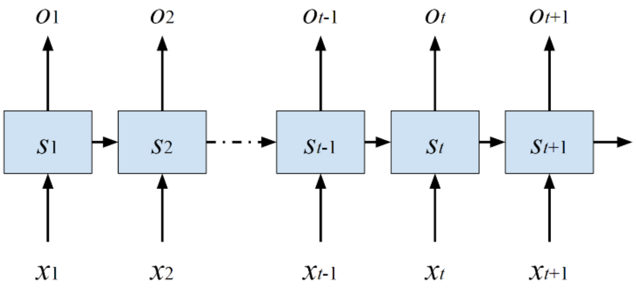


Figure 2: Recurrent neural network (RNN).

[4] http://www.speech.sri.com/projects/srilm/

Compared to n-gram, RNN based language model can capture longer distance dependencies with less parameters. As mentioned in Section 3, long distance dependency modelling is crucial to preposition selection. For this reason, we employ the simple RNN and the GRU language models to estimate the probability of a preposition in a given sentence. Instead of the single sigmoid or tanh function used by the simple RNN, GRU model, which is simplified from LSTM, uses a structure namely gated recurrent unit in the hidden layer.

$$
\begin{gathered}
s_t = (1 - z)h + zs_{t-1} \\
h = \tanh(W_h x_t + r_t(U_h s_{t-1})) \\
z = \sigma(W_z x_t + U_z s_{t-1}) \\
r = \sigma(W_r x_t + U_r s_{t-1})
\end{gathered}
$$

As shown in Figure 3, the update gate *z* decides how much the hidden state *s* is updated with the candidate hidden state $\tilde{s}$, while the reset gate *r* decides how much the memory to be forgotten. GRU is reported to be better for long-term dependency modeling (Chung et al., 2014; Chen et al., 2015b) than the simple RNN. Compared to LSTM, GRU requires few parameters for the same size of hidden layer.

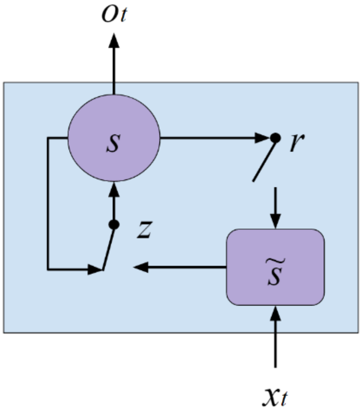


Figure 3: Gated recurrent unit (GRU).

In this work, we use the noise contrastive estimation (NCE) (Chen et al., 2015a) as the output layer for both simple RNN and GRU language models. The training performance of NCE is comparable to the class-based softmax, but NCE is faster in testing stage and can speed up together with GPU in training stage. The implementation of RNN models is based on faster-rnnlm, a toolkit for RNN language modelling[5].

6 Experiments

Three language models, n-gram, simple RNN, and GRU, are evaluated in the experiments with various configurations. The n-gram models are trained with the orders from 2 to 12. The simple RNN and the GRU models are trained with the hidden sizes of 128, 256, and 512. The number of noise samples of NCE is set to 20. Moreover, all language models are trained on three linguistic levels:

(1) Character level (**char**): each unit is a Chinese character.

(2) Word level (**word**): each unit is a Chinese word.

(3) Word level with POS tag (**w/p**): each unit is a combination of a word and its POS tag.

The performance is measured by accuracy and mean reciprocal rank (MRR). The accuracy is defined as number of correct selection versus number all test instances. The MRR is defined as $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$ where *N* is number of instances, and rank$_i$ is the rank of the correct preposition among *PS* according to its probability. In L1 testing, we check if the predicted preposition is the same as the original preposition in the sentence. In L2 testing, we check if the predicted preposition is the one corrected by annotators. McNemar test is used for significance testing at *p*=0.05.

[5] https://github.com/yandex/faster-rnnlm

Experimental results in L1 and L2 are shown in Table 2. For the n-gram models with different orders (from 2 to 12), we only show the highest performances due to the limited space. The subscripts of the simple RNN and the GRU models denote their hidden sizes. The best performing configuration for each of n-gram, simple RNN, and GRU models is highlighted in bold. In general, the larger the hidden layer, the better the performances of the simple RNN and the GRU language models. GRU model outperforms simple RNN and n-gram models in most cases. The best n-gram model is trained on word level with POS tags, the best simple RNN model is trained on word level using a hidden size of 512, and the best GRU model, which is also the best model of all, is trained on word-level with POS tag using a hidden size of 512. In both L1 and L2 testing, the best GRU model significantly outperforms the best simple RNN model.

The n-gram model and the GRU model perform better on the word level with POS tag, while the simple RNN model performs better on the word-level. On the other hand, all the n-gram, simple RNN, and GRU models trained on character level perform poorly. Grouping characters into words may help language models to capture long distance dependency. The neural networks are capable of learning feature representation from raw data. The linguistic information like POS tags still increases the performances of the GRU model.

The best performing model, GRU$_{512}$ trained on word level with POS tag, achieves an accuracy of 74.05% and an MRR of 83.08% on L1, and an accuracy of 60.13% and an MRR of 72.54% on L2. The precision@3 is 91% in L1 testing, and 81% in L2 testing. For a task of 43-way classification, this result is promising.

Among the three best performing n-gram, simple RNN and GRU models, the smallest performance gap between L1 and L2 is found in the GRU model. Compared to the other two models, GRU model achieves better testing fitness and less overfitting. However, the performance gap between L1 and L2 is still an issue to be tackled in the future.

LM	CGW (L1)						HSK (L2)					
	Character		Word		Word/POS		Character		Word		Word/POS	
	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR
N-gram	66.90	76.54	66.78	76.49	**69.11**	**78.68**	40.54	54.91	42.42	55.73	**43.62**	**56.65**
RNN$_{128}$	33.47	48.34	67.39	77.50	67.14	77.46	25.91	40.67	50.47	63.26	49.13	63.12
RNN$_{256}$	42.87	57.98	68.61	78.64	67.25	77.65	25.64	42.10	50.74	63.95	49.80	63.12
RNN$_{512}$	51.44	64.55	**71.80**	**81.03**	71.04	80.40	35.57	49.60	**55.97**	**68.56**	50.34	63.84
GRU$_{128}$	45.11	58.57	63.78	74.91	68.09	78.41	27.11	41.77	48.59	62.20	51.14	64.91
GRU$_{256}$	49.94	63.35	70.24	79.89	72.05	81.48	34.50	49.03	55.97	68.46	57.99	70.27
GRU$_{512}$	55.77	68.56	72.42	81.71	**74.05**	**83.08**	40.94	55.56	56.78	69.92	**60.13**	**72.54**

Table 2: Accuracies of n-gram, simple RNN, and GRU models with different configurations in L1 and L2. All the numbers are shown in percentage (%).

7 Discussion

Table 3 shows the performances of most frequent prepositions by the best performing n-gram, simple RNN and GRU models in L2 testing. The last row represents the performances of all prepositions in micro average. In each row, the best precision (P), recall (R), and F-score (F) are highlighted. The confusion matrix of the best performing GRU model in L2 testing is shown in Table 4. Most frequent prepositions are listed. Each row represents the sample in an actual preposition, while each column of the matrix represents the samples in a predicted preposition.

The preposition 在 (zài), the most frequent Chinese preposition, covers the meanings of *on*, *in*, and *at* in English. The precision of the second preposition 對 (duì, "to/at") achieved by the GRU model is only 59.76%. As shown in Table 4, 18 instances of 在 (zài, "on/in/at") and 10 instances of 給 (gěi, "for/to") are misclassified to 對. As a result, the recall of 給 (gěi, "for/to/with") and the precision of 對 are poor. In fact, they are sometimes interchangeable. For instance, using 對 in place of 給 in (S6) is also correct. However, only one correct preposition is labeled in the HSK corpus. In other words, our models are under-estimated due to the one-answer evaluation. In English, prepositions are also reportedly more than one way to correct (Bryant and Ng, 2015).

(S6) 吸菸 **給** 人 們 的 健康 帶來 不好 的 影響 (smoking causes bad effect **to** human health)

The third frequent preposition 從 (cóng, “from”) tends to confuse with 在 (zài, “on/in/at”) and 以 (yǐ, “by/with”). The fourth and the fifth prepositions 用 (yòng, “by/with”) and 以 (yǐ, “by/with”) share similar meanings and tend to be confusing.

Preposition	N-Gram Word/POS			RNN$_{512}$ Word			GRU$_{512}$ Word/POS		
	P	R	F	P	R	F	P	R	F
在 (on/in/at)	52.22	69.86	59.77	68.66	**84.02**	75.56	**79.70**	73.52	**76.48**
對 (to/at)	50.66	61.11	55.40	**63.70**	73.81	68.38	59.76	**80.16**	**68.47**
從 (from)	70.59	26.09	38.10	**87.18**	36.96	51.91	74.70	**67.39**	**70.86**
用 (by/with)	63.16	32.43	42.86	71.43	27.03	39.22	**77.78**	**37.84**	**50.91**
以 (by/with)	27.59	25.00	26.23	34.55	59.38	43.68	**41.51**	**68.75**	**51.76**
跟 (with)	**60.00**	19.35	29.27	50.00	**38.71**	**43.64**	**60.00**	29.03	39.13
由 (by/from)	33.33	31.82	32.56	54.55	54.55	54.55	**72.22**	**59.09**	**65.00**
給 (for/to/with)	50.00	5.88	10.53	**100.00**	5.88	11.11	**100.00**	**17.65**	**30.00**
Average of all	46.14	43.62	41.71	60.15	56.78	55.88	**63.63**	**60.13**	**59.20**

Table 3: Performances of most frequent prepositions by the best performing n-gram, simple RNN and GRU models in L2 testing. All numbers are shown in percentage (%).

Actual	Predicted Prepositions							
	在 (on/in/at)	對 (to/at)	從 (from)	用 (by/with)	以 (by/with)	跟 (with)	由 (by/from)	給 (for/to/with)
在 (on/in/at)	161	18	7	1	1	1	0	0
對 (to/at)	8	101	3	0	4	1	1	0
從 (from)	10	3	62	1	9	1	0	0
用 (by/with)	2	3	2	14	9	0	0	0
以 (by/with)	0	4	2	0	22	1	0	0
跟 (with)	2	3	0	0	0	9	0	0
由 (by/from)	1	1	0	2	1	0	13	0
給 (for/to/with)	0	10	0	0	0	0	0	3

Table 4: Confusion matrix of the GRU$_{512}$ on the word level with POS tag in L2 testing.

Figure 4 shows the accuracies of the n-gram, RNN$_{512}$, and GRU$_{512}$ models on word level with and without POS tag with respect to difference sentence lengths. We divide the sentences in L2 test set into five groups according to their length. The average sentence length is 9.13 words, and the longest sentence consists of 32 words. The information of POS tag generally improves the performance for the cases of longer sentences. In particular, the GRU$_{512}$ model on word-level with POS tag outperforms other models for the sentences of lengths ≥ 4 words. In the test set of L2, only 10 sentences (1.3%) are shorter than 4 words.

(S7) is an instance correctly predicted by the GRU$_{512}$ model, while n-gram and RNN$_{512}$ output a wrong outcome as (S8). In this case, the first word 在 (zài, “in”) and the last word 裡 (lǐ, “inside”) form a circumposition for the noun phrase 那些保守的家庭 (those conservative family). The pair of words 在... 裡 (inside) is a common usage in Chinese. The GRU$_{512}$ model successfully captures their dependency although there are four words in-between.

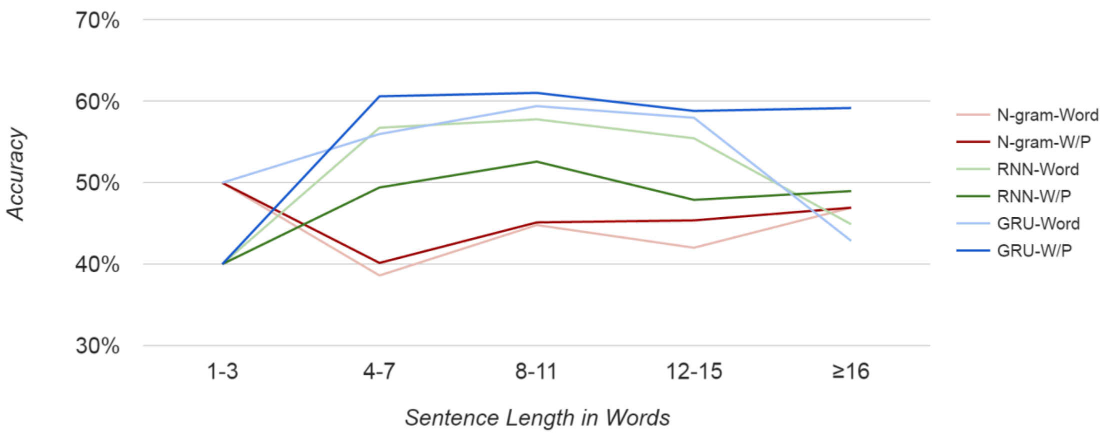


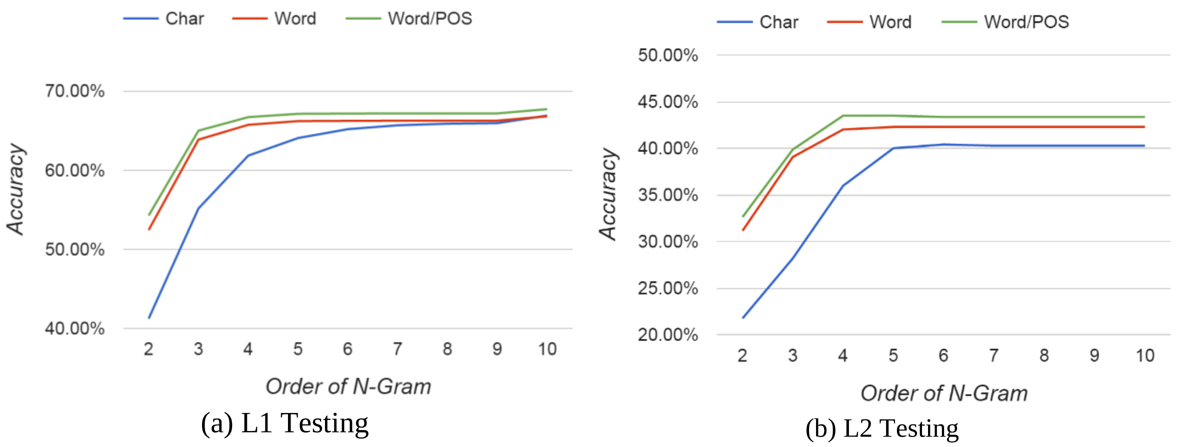
Figure 4: Accuracy versus sentence length in L2 testing.

```
(S7) 在 那些 保守 的 家庭 裡 (Inside those conservative family)
(S8) 對 那些 保守 的 家庭 裡 (To those conservative family)
```

(S9) shows a much longer distance dependency successfully estimated by GRU512. In this case, there are eight words in-between the word pair 以...為 (aim at ... as). The RNN512 selects 就 (jiù, “just/on”) in the sentence (S10), which is not fluent. The n-gram model, even worse, selects 在 (zài, “in”) and makes an incomprehensible sentence (S11).

```
(S9) 一些 家長 也 以 把 孩子 送入 純 女 或 純 男校 為 第一 選擇 (Some
parents aim at sending their children into the pure female or
pure male school as the first choice)
(S10) 一些 家長 也 就 把 孩子 送入 純 女 或 純 男校 為 第一 選擇
(S11) 一些 家長 也 在 把 孩子 送入 純 女 或 純 男校 為 第一 選擇
```

Figure 5 illustrates accuracies of varying order of the n-gram on character level, word level, and word level with POS tag. Results of L1 and L2 testing are shown in Figure 5(a) and Figure 5(b), respectively. As the order of n-gram increases, the models on word level and on word level with POS tag faster growth. In L1 testing, the n-gram model on character level finally achieves an accuracy close those of other two models on word level. In L2 testing, the performance gaps among the three models are more apparent. The information of Chinese word segmentation and POS tags not only speeds up training, but also improves the generalization.



(a) L1 Testing

(b) L2 Testing

Figure 5: Accuracy versus order of n-gram in L1 and L2 testing.

Previous work suggests that the ensemble of RNN and traditional language model may improve the performance, especially for the RNN models with small hidden layer (Mikolov, 2012; Mikolov et al., 2011). We build an ensemble model GRU-ME by joint training the GRU model with a 4-gram maximum entropy language model (Berger et al., 1996). The input unit is word with POS tag (w/p), which has best performance in the experiments. The feature size of maximum entropy model is 1 billion. Figure 6 compares the performances between GRU and GRU-ME in L1 and L2. In L1 testing, ensembling the maximum entropy model with GRU significantly increases the performances of three GRU models, especially the ones with smaller hidden layer. In L2 testing, ensembling the maximum entropy model increases the performances of the GRU model with hidden sizes of 128. The results confirm that the RNN models with smaller hidden size gain from the combination of the traditional language model. The performances of the GRU models with larger hidden layer, however, are decreased with the combination of maximum entropy model and GRU. This phenomenon suggests that GRU-ME better fits the training data, but may suffer from overfitting. In contrast, the GRU model with a large hidden size is more robust when it is applied to another corpus.

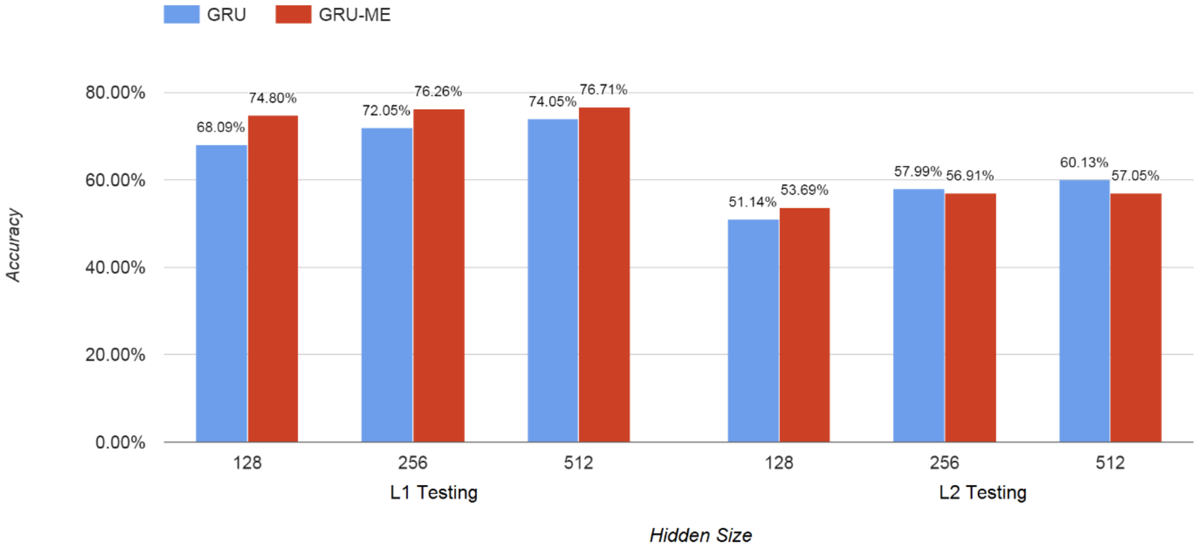


Figure 6: Performances of the GRU and the ensemble of GRU with Maxent (GRU-ME) models in L1 and L2 testing.

8 Conclusion

This work addresses the issue of Chinese preposition selection. We propose a method that uses language models to predict the most probable preposition in a given context. The classical n-gram models and the recurrent neural network models are explored. For the task of modelling Chinese prepositions, the experimental results show the advantage of the GRU models over simple RNN and n-gram models, especially for the cases involving longer distance dependency. In addition, linguistic information from Chinese word segmentation and the POS tagging improve the performances of n-gram and neural network language models. We will further adapt this approach to detection and correction for other grammatical errors in future work.

9 Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-104-2221-E-002-061-MY3 and MOST-105-2221-E-002-154-MY3, and National Taiwan University under grant NTU-ERP-104R890858.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):1-36.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2009. Web-Scale N-gram Models for Lexical Disambiguation. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence* (*IJCAI-09*), pages 1507-1512.

Christopher Bryant and Hwee Tou Ng. 2015. How Far are We from Fully Automatic High Quality Grammatical Error Correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP*), pages 697–707, Beijing, China.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust Systems for Preposition Error Correction Using Wikipedia Revisions. In *Proceedings of NAACL-HLT 2013*, pages 507–517.

Xie Chen, Xunying Liu, Mark J. F. Gales, and Philip C. Woodland. 2015a. Recurrent Neural Network Language Model Training with Noise Contrastive Estimation for Speech Recognition. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5411-5415, South Brisbane, Australia.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuanjing Huang. 2015b. Gated Recursive Neural Network for Chinese Word Segmentation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1744–1753, Beijing, China.

Shuk-Man Cheng, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. In *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING*), pages 23-29, August 2014, Dublin, Ireland.

Chinese Knowledge Information Processing Group (CKIP). 1993. Technical Report 93-05: Chinese Part-of-Speech Analysis. Academia Sinica, Taipei.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078.

Martin Chodorow, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of Grammatical Errors Involving Prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv preprint arXiv:1412.3555v1

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 54–62, Montreal, Canada.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, Nancy, France.

Rachele De Felice and Stephen G. Pulman. 2007. Automatically Acquiring models of Preposition Use. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 45–50, Prague, Czech Republic.

Rachele De Felice and Stephen G. Pulman. 2008. A Classifier-based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pages 169–176.

Chu-Ren Huang, Lung-Hao Lee, Jia-Fei Hog, Wei-guang Qu, and Shiwen Yu. 2008. Quality Assurance of Automatic Annotation of Very Large Corpora: a Study based on heterogeneous Tagging System. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*), pages 2725–2729.

Chu-Ren, Huang. 2009. *Tagged Chinese Gigaword Version 2.0 LDC2009T14*. Web Download. Philadelphia: Linguistic Data Consortium.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang. 2015a. Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (*NLP-TEA*), pages 1–6, Beijing, China.

Lung-Hao Lee, Gina-Anne Levow, Shih-Hung Wu, and Chao-Lin Liu. 2015b. Introduction to the Special Issue on Chinese Spell Checking. *ACM Transactions on Asian and Low-Resource Language Information Processing* (*Special Issue on Chinese Spell Checking*), 14(4):14.

Jia-Na Lin. 2011. *Analysis on the Biased Errors of Word Order in Written Expression of Foreign Students*. Master Thesis. Soochow University.

Chuan-Jie Lin and Shao-Heng Chan. 2014. Description of NTOU Chinese Grammar Checker in CFL 2014. In *Proceedings of the 22nd International Conference on Computers in Education*, pages 75–78, Nara, Japan.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. *Penn Treebank 3 LDC99T42.* Web Download. Philadelphia: Linguistic Data Consortium.

Tomas Mikolov. 2012. *Statistical Language Models Based on Neural Networks.* Doctoral Dissertation. Brno University of Technology.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukas Burget, and Jan "Honza" Cernocky. 2011. RNNLM – Recurrent Neural Network Language Modeling Toolkit. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop.*

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task,* pages 1–14, Baltimore, Maryland.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining Geometric, Textual and Visual Features for Predicting Prepositions in Image Descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 214–220, Lisbon, Portugal.

Yow-Ting Shiue and Hsin-Hsi Chen. 2016. Detecting Word Usage Errors in Chinese Sentences for Learning Chinese as a Foreign Language. In *Proceedings of 10th Language Resources and Evaluation Conference*, pages 220-224, Portorož, Slovenia.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using Parse Features for Preposition Selection and Error Detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358, Uppsala, Sweden.

Zhuo Wang. 2011. *A Study on the Teaching of Unique Syntactic Pattern in Modern Chinese for Native English Speaking Students*. Master Thesis. Northeast Normal University.

Yang Xiang, Bo Yuan, Yaoyun Zhang, Xiaolong Wang, Wen Zheng, and Chongqiang Wei. 2013. A Hybrid Model for Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning* (*CONLL*): *Shared Task*, pages 115–122, Sofia, Bulgaria.

Nianwen Xue, Xiuhong Zhang, Zixin Jiang, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. 2013. *Chinese Treebank 8.0 LDC2013T21*. Web Download. Philadelphia: Linguistic Data Consortium.

Chi-Hsin Yu and Hsin-Hsi Chen. 2012. Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language. In *Proceedings of COLING 2012: Technical Papers*, pages 3003–3018, COLING 2012, Mumbai, India.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. In *Proceedings of the 22nd International Conference on Computers in Education*, pages 42-47.

Marcos Zampieri and Liling Tan. 2014. Grammatical Error Detection with Limited Training Data: The Case of Chinese. In In *Proceedings of the 22nd International Conference on Computers in Education*, pages 69-74, Nara, Japan.

Longkai Zhang and Houfeng Wang. 2014. A Unified Framework for Grammar Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 96–102, Baltimore, Maryland.

Yinchen Zhao, Mamoru Komachi, and Hiroshi Ishikawa. 2014. Extracting a Chinese Learner Corpus from the Web: Grammatical Error Correction for Learning Chinese as a Foreign Language with Statistical Machine

Translation. In *Proceedings of the 22nd International Conference on Computers in Education*, pages 56–61, Nara, Japan.

Yinchen Zhao, Mamoru Komachi, and Hiroshi Ishikawa. 2015. Improving Chinese Grammatical Error Correction using Corpus Augmentation and Hierarchical Phrase-based Statistical Machine Translation. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 111–116, Beijing, China.