

Word Order Sensitive Embedding Features/Conditional Random Field-based Chinese Grammatical Error Detection

Wei-Chieh Chou, Chin-Kui Lin,
Yih-Ru Wang
Department of Electrical Engineering,
National Chiao Tung University, Hsinchu,
Taiwan, ROC
m0450743.eed04g@g2.nctu.edu.tw, lck199382.eed04g@g2.nctu.edu.tw, yrwang@mail.nctu.edu.tw

Yuan-Fu Liao
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan, ROC
yfliao@ntut.edu.tw

Abstract

This paper discusses how to adapt two new word embedding features to build a more efficient Chinese Grammatical Error Diagnosis (CGED) systems to assist Chinese foreign learners (CFLs) in improving their written essays. The major idea is to apply word order sensitive Word2Vec approaches including (1) structured skip-gram and (2) continuous window (CWindow) models, because they are more suitable for solving syntax-based problems. The proposed new features were evaluated on the Test of Chinese as a Foreign Language (TOCFL) learner database provided by NLP-TEA-3&CGED shared task. Experimental results showed that the new features did work better than the traditional word order insensitive Word2Vec approaches. Moreover, according to the official evaluation results, our system achieved the lowest (0.1362) false positive (FA) and the highest precision rates in all three measurements among all participants.

1 Introduction

In recent years, the rise of Asian economies and nearly 20 years of rapid development of China has led to a corresponding interest in the study of Standard Chinese ("Mandarin") as a foreign language, the official language of mainland China and Taiwan.

However, it might be a great challenge for those CFLs to learn how to write an essay or report in Chinese. Because approximately 3,000 Chinese characters and 5,000 words are required for receiving Test of Chinese as a Foreign Language (TOCFL) certificate in advanced level¹. Besides, Chinese is an analytic language, in that they depend on syntax (word order and sentence structure) rather than morphology, i.e., changes in form of a word, to indicate the word's function in a sentence. And Chinese also makes heavy use of grammatical particles to indicate aspect and mood, such as like 了 (le, perfective), 還 (hái, still), 已經 (yǐjīng, already), and so on.

CFLs often make four types of grammatical errors, including (1) disorder, (2) missing, (3) redundant and (4) selection, for example:

- Disorder: 我要送給你一個慶祝禮物。要是兩、三天晚了，請別生氣 ("兩、三天晚了" should be "晚了兩、三天")
- Missing: 我聽說你找到工作。恭喜恭喜! ("工作" should be "工作了")
- Redundant: 今天是我大學畢業了 ("今天是" should be "今天")
- Selection: 我等在教室沒那麼久老師就來了 ("那麼" should be "多")

To detect those grammatical errors is not an easy task. Recently, researchers have proposed many approaches for CGED task. They could be roughly divided into two categories including (1) hybrid linguistic rules+language modelling and (2) pure classification-based methods.

¹ http://www.sc-top.org.tw/english/eng_index.php

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

For example, Lee et al. (2013) applied a set of handcrafted linguistic rules with syntactic information to detect errors occurred in Chinese sentences written by CFLs. Lee et al. (2014) then further implemented a sentence judgement system that integrated both rule-based and n-gram statistical methods to detect grammatical errors in Chinese sentences. Lin and Chen (2015) proposed a system which measured the likelihood of sentences generated by deleting, inserting, or exchanging characters or words in which two sentence likelihood functions were proposed based on frequencies of space removed version of Google n-grams.

On the other hand, Xiang (2015) utilized an ensemble classifier random feature subspace method for CGED task. Cheng et al. (2014) proposed a CRF-based method to detect word ordering errors and a ranking SVM-based model to suggest the proper corrections. Finally, Chen et al. (2015) and Yeh et al. (2015) also adopt CRFs and collected a set of common grammatical error rules for building CGED systems.

Among these two methods, the classification-based approach, especially the CRF-based one is quite promising. Because, CRFs treat the CGED problem as a sequence-to-sequence mapping task, it could then model well the word ordering and sentence structure. However, traditional CRF-based approaches often only take current and few neighbouring words and their POS tags as the input features. This may limit CRFs' horizon vision. Besides, word-based features will result in the sparse training data problem, since the total number of Chinese words is more than 160,000².

To alleviate these two difficulties, this paper would like to discuss how to adapt word embedding features to alleviate the sparseness issue and especially how to extract two new word order sensitive embedding features proposed by Wang (2015) to capture ordering information. The major idea is to apply word order sensitive Word2Vec approaches including (1) Structured Skip-gram and (2) CWindow models. Because they seriously take word ordering information into account and are therefore more suitable for solving syntax-based problems. By this way, we hope we could build a more efficient CGED system.

2 System Implementation for NLP-TEA-3&CGED shared task

The block diagram of our proposed system is shown in Fig. 1. It has a CRF-based traditional Chinese parser for word segmentation and POS tagging frontend and a CRF-based CGED backend. But the major enhancement comparing with other CRF-based approaches is that it applies the word order sensitive Word2Vec module to extract word embedding vectors and then does word clustering to generate input features for CRF-based CGED module.

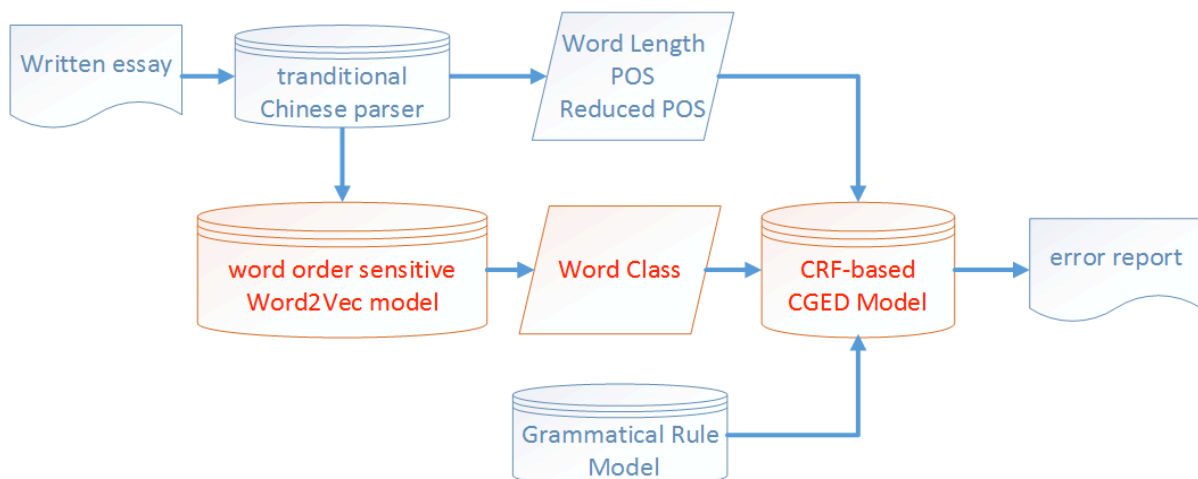


Figure 1: The block diagram of the proposed word ordering sensitive embedding feature/CRF-based Chinese Grammatical Error Detection system.

² <https://www.moedict.tw/about.html>

In the following subsections, several system components will be discussed in more detail, including (1) traditional Chinese parser, (2) word order sensitive Word2Vec, (3) grammatical rule and (4) CRF-based CGED models.

2.1 Traditional Chinese Parser

The parser used in this system (as shown in Fig. 2) is a CRF-based system for traditional Chinese. It has three main modules including (1) text normalization, (2) word segmentation and (3) POS tagging.

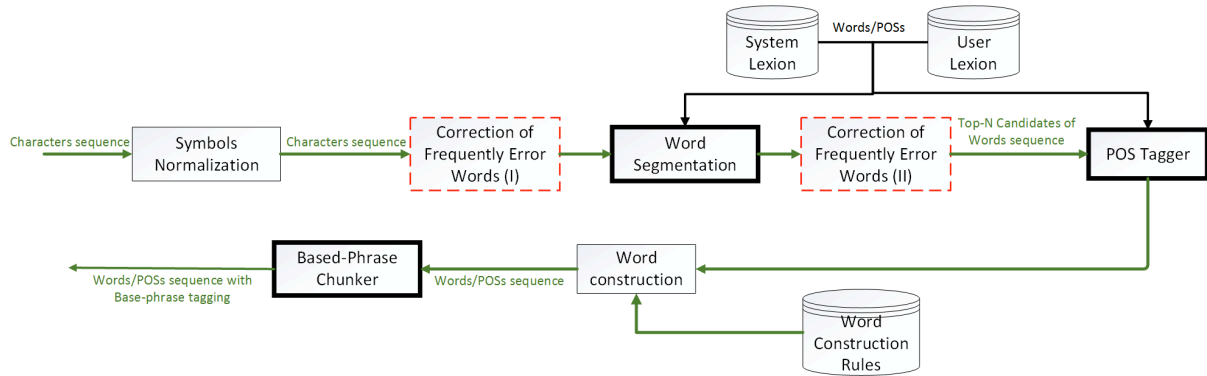


Figure 2: The schematic diagram of the proposed Chinese parser.

This parser was trained using Sinica Balanced Corpus ver. 4.0³. Its performance is as follows: The F-measure of the word segmentation is 96.72% for the original database and 97.50% for the manually corrected corpus. The difference between precision and recall rate is less than 0.06%. The accuracy of the POS (47-type) tagging is 94.97%.

It is worth noting that this CRF-based Chinese word segmentation and parser is originally built for automatic speech recognition (ASR). So another purpose of this study is to examine how generalization and sophistication our parser is. Since Chinese words are not well defined (without word boundaries), a high quality Chinese word segmentation and parser is essential for building an effective word embedding representation and a good CGED system.

2.2 Word Order Sensitive Embedding Feature Extraction

One way to alleviate the sparse training data problem is to use word classes instead of words themselves as the input features for CRF-based CGED system. The most widely used tools for building word clustering are the models described in (Mikolov 2013a, b, c), including the “Skip-gram” and the “Continuous Bag-of-Words” (CBOW) models. However, since these models only give a word a single embedding feature vector, they are insensitive to word order and may not be suitable for CGED tasks.

Therefore, in this paper, we will adopt two new word order sensitive embedding approaches including (1) CWindow and (2) Structured Skip-gram (see Fig. 3) models proposed by Wang (2015) to take word ordering information into account.

Basically, CWindow defines a output predictor $O \in \mathbb{R} (|V| \times (2c \times d))$ that takes a $(2c \times d)$ -dimensional vector $[e(w_{-c}), \dots, e(w_{-1}), e(w_1), \dots, e(w_c)]$ (the embeddings of the context words) as input. Words in different position hence have different weights. Structured Skip-gram, on the other hand, defines a set of $c \times 2$ output predictors ($O_{-c}, \dots, O_{-1}, O_1, O_c$), with size $O \in \mathbb{R} <(|V|) \times d$, to predict the outputs according to their positions. These two models then will generate word order sensitive embeddings features. By this way, it should be easier for CGED system to detect abnormal word ordering or sentence structures.

³ http://www.aclclp.org.tw/use_acbc_c.php

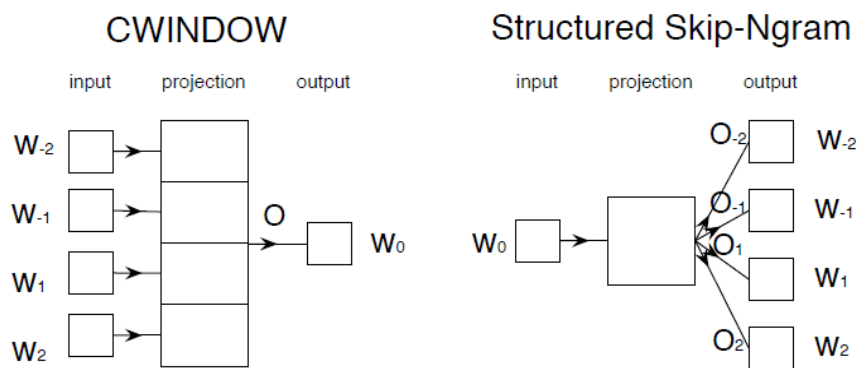


Figure 3: Illustration of the CWindow and Structured Skip-gram models.

2.3 Error-Prone Words

To detect some common grammatical errors often made by CFLs, a set of error patterns could be used to assist CGED system. For example, Yeh (2015) checked the following words in his CRF-based CGED system:

- Quantifiers
 - ✓ #. of human: “位” or “個”
 - ✓ #. of animals: “隻”, “匹”, “頭”, or “條”
 - ✓ #. of things: “件”
 - ✓ #. of buildings: “座” or “棟”
 - ✓ #. of vehicle: “臺”, “輛”, “架” or “艘”
- Error-prone words (word + POS)
 - ✓ “把 (let)” + “Nh” or “Na” or “Nep”
 - ✓ “跟(with)” + “VA” or “Nh” or “Na”
 - ✓ “應該(maybe)” or “好像(like)” or “到底(at last)” + “Nh” or “Na”
 - ✓ “已經(already)” + “Neqa” or “Neu”
 - ✓ “Neqa” or “Neu” + “P” or “Na” or “VA”

After some literature survey, it is found that CFLs also often use “是(is)”, “的(of)” and “了(finish)” incorrectly. Therefore, this work will specially considers a set of 9 error-prone words including “把(let)”, “跟(with)”, “應該(maybe)”, “好像(like)”, “到底(in the end)”, “已經(already)”, “是(is)”, “的(of)” and “了(finish)”. Here are some real error examples produced by CFLs (from TOCFL learner corpus):

- Redundant: “把(let)”, “是(is)” or “的(of)” or “了”
 - ✓ 看著我把搖搖頭。(“看著我把搖搖頭” should be “看著我搖搖頭”)
 - ✓ 她是很漂亮。(“她是很漂亮” should be “她很漂亮”)
 - ✓ 張愛文住在台北，他是十六歲。(“他是十六歲” should be “他十六歲”)
 - ✓ 我想請你在日本料理的餐廳吃飯，好不好？(“我想請你在日本料理的餐廳吃飯” should be “我想請你在日本料理餐廳吃飯”)
 - ✓ 幸虧沒有別的人先買走了。(“幸虧沒有別的人先買走了” should be “幸虧沒有別的人先買走”)
- Disorder: “跟(with)”, “應該(should)”, “好像(seem)”, “到底(in the end)”
 - ✓ 要不要下了課去西門跟我看電影？(“要不要下了課去西門跟我看電影” should be “要不要下了課跟我去西門看電影”)
 - ✓ 家長也應該不讓小孩子亂玩網路遊戲。(“家長也應該不讓小孩子亂玩網路遊戲” should be “家長也不應該讓小孩子亂玩網路遊戲”)
 - ✓ 到底我哪裡有錯？(“到底我哪裡有錯” should be “我到底哪裡有錯”)

- ✓ 現在我跟這瑞典人的關係是已經很好的朋友。（“現在我跟這瑞典人的關係是已經很好的朋友” should be “現在我跟這瑞典人的關係已經是很好的朋友”）
- Selection: “好像”
 - ✓ 我覺得為了知道他們的想法，學生應該學他們的文化。好像他們的生活方式，歷史甚麼的。（“好像他們的生活方式” should be “例如他們的生活方式”）
- Missing: “了(finish)”
 - ✓ 你找到工作，恭喜恭喜！（“你找到工作” should be “你找到工作了”）

2.4 CRF-based Chinese Grammatical Error Detection

The performance of CRF-based model is mainly decided by the quality of feature engineering. In this work, five different features and four sets of feature templates are designed for building our CRF-based CGED system.

First, the five features are (1) word length, (2) POS, (3) reduced POS⁴, (4) word class index, (5) error-prone word indicator. Fig. 4 shows an example of how these features are putting together with the grammatical error-type ground-truth to form a training data file.

你	1	Nh	N	84	0	0
開	1	VC	Vt	546	0	0
一	1	Neu	DET	374	0	B-R
個	1	Nf	M	624	0	I-R
慶祝會	3	Na	N	834	0	0
的	1	DE2	T	752	的	0
時候	2	Na	N	126	0	0
我	1	Nh	N	84	0	0
不能	2	D	ADV	715	0	0
會	1	D	ADV	834	0	B-R
參加	2	VC	Vt	952	0	0
，	1	PM	PM	-2	0	0
是	1	SHI	Vt	622	是	0
因為	2	Cbb	C	622	0	0
我	1	Nh	N	84	0	0
在	1	P	P	116	0	0
外國	2	Nc	N	6	0	0
做	1	VC	Vt	546	0	B-R
工作	2	Na	N	81	0	0

Figure 4: An example list of features and error-type ground-truth for training our CRF-based CGED model. The columns from left to right are word, word length, POS, reduced POS, word class index, error-prone word indicator and grammatical error-type ground-truth, respectively.

Second, the four sets of feature templates are specified in Table 1. They consider some combinations of the five features (and their n-grams). It is worth noting that the fourth template “(POS_{n-1} EPWI_n POS_{n+1})” could be treated as a generalization of the “Error-prone words (word + POS)” pattern proposed in Yeh (2015) (mentioned in previous subsection).

Features	Features Templates
Word Class	$C_{n-2}, C_{n-1}, C_n, C_{n+1}, C_{n+2}, (C_{n-2} C_{n-1} C_n), (C_n C_{n+1} C_{n+2}), (C_{n-1} C_n C_{n+1}), (C_{n-2} C_{n-1} C_n C_{n+1} C_{n+2})$
POS+RPOS	$(RPOS_{n-2} RPOS_{n-1} POS_n), (RPOS_{n-1} POS_n RPOS_{n+1}), (POS_n RPOS_{n+1} RPOS_{n+2})$
RPOS+Word Class	$(RPOS_{n-2} RPOS_{n-1} C_n), (RPOS_{n-1} C_n RPOS_{n+1}), (C_n RPOS_{n+1} RPOS_{n+2})$
Error-Prone Word Indicator	$(POS_{n-1} EPWI_n POS_{n+1})$

Table 1: List of feature templates designed for building our CRF-based CGED system. Here “C”, “POS”, RPOS and EPWI are the word class index, POS, reduced POS and error-prone word indicator, respectively.

⁴ <http://ckipsvr.iis.sinica.edu.tw/cat.htm>

3 NLP-TEA 3 & CGED Shared Task

The goal of the NLP-TEA3&CGED shared task is to develop systems to automatically diagnose Chinese sentences written by CFLs. The systems should indicate where and what type of errors are embedded in CFLs' sentence.

In the following experiments, the effectiveness of the error-prone word templates was first checked. Then the performance of the new "CWindow" and "Structured Skip-gram" were compared with the original "Skip-gram" and the "CBOW" models. Finally, the official evaluation results of our three CWindow-based submissions were discussed.

3.1 TOCFL learner database

The TOCFL learner database (NLP-TEA3) provided by the organizers was used to develop our CGED system. In order to enlarge the pool of training samples, the data sets of the two previous editions of this shared task, i.e., NLP-TEA1 (Yu et al. (2014)) and NLP-TEA2 (Lee et al. (2015)) are also added together. In the end, there are in total 63,462 sentences for system development. Table 2 shows the statistics of different grammatical error types on the development dataset.

Error-type	#. of errors
Disorder	1,980
Redundant	4,971
Missing	90
Selection	10,686
Correct	35,141

Table 2: Statistics of the numbers of error-types made by CFLs on our training corpus.

The development data was further divided into a training and a testing subsets by a ratio of 9:1. Therefore, there are 57,116 and 6,346 sentences in the training and testing subsets, respectively.

3.2 Model Settings

Four types of embedding representations including CBOW, Skip-gram, Structured Skip-gram and CWindow models were built using the modified Word2Vec toolkit^{5,6}. They were all trained using the same set of text corpora including (1) LDC Chinese Gigaword Second Edition⁷, (2) Sinica Balanced Corpus ver. 4.0, (3) CIRB0303⁸ (Chinese Information Retrieval Benchmark, version 3.03), (4) Taiwan Panorama Magazine⁹, (5) TCC300¹⁰ and (6) Wikipedia (ZH_TW version).

In all methods, the vector size was set to 300 and using a context window of 13 (6+1+6) words. To speed up the computation, the probability of a target word was estimated with the hierarchy Softmax method. After the vector space is established, k-mean algorithm was utilized to cluster all words into 1,024 classes.

Finally, the CRF++ toolkit developed by Kudo¹¹ was utilized to build our CRF-based CGED system. It is worth noting although there are four different Word2Vec frontends but the CRF backend is the same (except the input word class features) for all following experiments.

3.3 Preliminary Results on Development Dataset

First of all, Table 3 shows the impact of (with and without) error-prone words evaluated using a CWindow/CRF-based system. According the results, it indicates that those special words did help to

⁵<https://code.google.com/p/word2vec/>

⁶<https://github.com/dav/word2vec>

⁷<https://catalog.ldc.upenn.edu/LDC2005T14>

⁸http://www.aclclp.org.tw/use_cir.php

⁹http://www.aclclp.org.tw/use_gh_c.php

¹⁰http://www.aclclp.org.tw/use_mat.php - tcc300edu

¹¹<https://taku910.github.io/crfpp/>

improve the performance of our CRF-based CGED system. Therefore, those error-prone words will be considered in all systems reported below.

Approach	Accuracy	Precision	Recall	F1
Without	89.91%	52.17%	10.69%	17.75%
With	89.89%	52.32%	10.89%	18.03%

Table 3: Performance comparison on the effectiveness of adding the error-prone words feature templates on a CWindow/CRF-based CGED system.

Second, Table 4 showed the performance of the Structured Skip-gram-, CWindow-, Skip-gram- and CBOW-based CGED systems (all take error-prone words into account). It is found that CWindow achieved the best F1-score. Because F1-score is the most balanced performance measurement, all our submissions will use the CWindow-based approach.

Approach	Accuracy	Precision	Recall	F1
Skip-gram	89.59%	44.57%	10.49%	16.98
CBOW	89.91%	52.00%	10.32%	17.22
Structured Skip-gram	90.02%	56.97%	10.19%	17.29
CWindow	89.89%	52.32%	10.89%	18.03

Table 4: Performance of the Structured Skip-gram-, CWindow-, Skip-gram- and CBOW-based CGED systems (all take error-prone words into account).

3.4 Official Evaluation Results

Three runs (NCTU+NTUT-Run1, Run2 and Run3) were submitted to NLP-TEA 2016 CGED shared task for official evaluation. All submissions are CWindow-based systems, since CWindow achieved the best performance in preliminary experiments. The only difference between these three runs is that they have different FA performance (i.e., different operating points). Table 5 shows the official evaluation results of our three submissions.

Among three submissions, Run1 has the lowest FA and highest precision rate in all three measurements comparing with other participants. Especially, Run1 achieved 0.1362 FA, 0.4603 accuracy, 0.2542 precision and 0.0483 recall rate in position-level. Since FA is the most important factor that influences users' experiences on CGED applications, the proposed approach is quite promising.

Team-Run	False Positive Rate	Detection-Level				Identification-Level				Position-Level			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
NCYU-Run1	0.5602	0.5507	0.5559	0.6542	0.6011	0.3577	0.2749	0.2862	0.2805	0.1728	0.0074	0.0056	0.0064
NCYU-Run2	0.9612	0.5218	0.5202	0.9726	0.6779	0.2328	0.2265	0.4744	0.3066	0.0231	0.0129	0.0195	0.0155
NCYU-Run3	0.8491	0.5363	0.5307	0.8959	0.6665	0.2653	0.2384	0.4134	0.3024	0.058	0.013	0.0163	0.0145
CYUT-Run1	0.347	0.5955	0.6259	0.5419	0.5809	0.5154	0.46	0.3021	0.3647	0.3113	0.1461	0.1089	0.1248
CYUT-Run2	0.3558	0.5955	0.6236	0.5501	0.5846	0.5133	0.4567	0.3061	0.3666	0.3061	0.1432	0.1092	0.1239
CYUT-Run3	0.3635	0.5941	0.6205	0.5545	0.5856	0.5078	0.4472	0.3001	0.3592	0.3088	0.1196	0.0768	0.0935
PKU-Run1	0.2284	0.521	0.5739	0.2871	0.3828	0.4575	0.3418	0.1173	0.1747	0.3844	0.0996	0.0263	0.0416
PKU-Run2	0.7205	0.5258	0.5292	0.7556	0.6224	0.3242	0.2792	0.3712	0.3187	0.1381	0.068	0.0824	0.0745
PKU-Run3	0.525	0.5349	0.5467	0.5907	0.5678	0.3705	0.2729	0.2192	0.2431	0.2331	0.0872	0.0651	0.0745
NCTU+NTUT-Run1	0.1362	0.5442	0.6593	0.246	0.3583	0.511	0.4892	0.1224	0.1958	0.4603	0.2542	0.0483	0.0811
NCTU+NTUT-Run2	0.2913	0.553	0.6	0.4077	0.4855	0.4793	0.4036	0.1982	0.2659	0.3784	0.1644	0.0639	0.092
NCTU+NTUT-Run3	0.32	0.5612	0.6013	0.4504	0.515	0.4773	0.3993	0.2185	0.2824	0.3613	0.1521	0.0668	0.0928
YNU-HPCC-Run1	0.6289	0.542	0.5444	0.7014	0.613	0.2211	0.1588	0.3196	0.2122	0.0886	0.0002	0.0002	0.0002
YNU-HPCC-Run2	0.5931	0.5026	0.5167	0.5918	0.5517	0.2322	0.1675	0.3136	0.2184	0.0991	0	0	null
YNU-HPCC-Run3	0.3382	0.4847	0.503	0.3195	0.3908	0.4023	0.281	0.1359	0.1832	0.2797	0.0012	0.0005	0.0007

Table 5: Official TOCFL evaluation results of NLP-TEA3&CGED shared task.

4 Conclusion

In this paper, a word order sensitive embedding features/CRF-based CGED system was proposed and implemented for participating the NLP-TEA-3&CGED shared task. The experimental results showed that the proposed new features did work better than the traditional word order insensitive Word2Vec approaches. Moreover, according to the official evaluation results, our system achieved the lowest FA (0.1342) and the highest precision rates in all three measurements among all participants. Therefore, the proposed approach is a promising one and will be further explored in the near future. Finally, the latest version of our traditional Chinese parser is available on-line at <http://parser.speech.cm.nctu.edu.tw>.

Acknowledgment

This work was supported by the Ministry of Science and Technology, Taiwan with contract 105-2221-E-009-142-MY2, 104-2221-E-027-079 and 105-2221-E-027-119

Reference

- Chen, Po-Lin, Wu, Shih-Hung. (2015). Chinese Grammatical Error Diagnosis by Conditional Random Fields. Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 7–14.
- Cheng, Shuk-Man, Yu, Chi-Hsin, Chen, Hsin-Hsi. (2014) Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.
- Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of ICML, pp.282-289, 2001
- Lee, Lung-Hao, Chang, Li-Ping, Lee, Kuei-Ching, Tseng, Yuen-Hsien, and Chen, Hsin-Hsi (2013). Linguistic Rules Based Chinese Error Detection for Second Language Learning. In Work-in-Progress Poster Proceedings of the 21st International Conference on Computers in Education (ICCE'13), Denpasar Bali, Indonesia, 18-22 November, 2013, pp. 27-29.
- Lee, Lung-Hao, Yu, Liang-Chih, Lee, Kuei-Ching, Tseng, Yuen-Hsien, Chang, Li-Ping, and Chen, Hsin-Hsi. (2014). A Sentence Judgment System for Grammatical Error Detection. In Proceedings of the 25th International Conference on Computational Linguistics (COLING'14), Dublin, Ireland, 23-29 August, 2014, pp. 67-70.
- Lee, Lung-Hao, Liang-Chih Yu, and Li-Ping Chang. 2015. 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 2015). 1-6.
- Lin, Chuan-Jie, and Chen, Shao-Heng. (2015). NTOU Chinese Grammar Checker for CGED Shared Task. Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 15–19,
- Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. (2013a). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.
- Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS.
- Mikolov, Tomas, Yih, Wen-tau, and Zweig, Geoffrey. (2013c). Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT.
- Wang, Ling, Dyer, Chris, Black, Alan, and Trancoso, Isabel, (2015). Two/Too Simple Adaptations of Word2Vec for Syntax Problems. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
- Wang, Yih-Ru, and Liao, Yuan-Fu (2015). Word Vector/Conditional Random Field-based Chinese Spelling Error Detection for SIGHAN-2015 Evaluation. Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 46–49, Beijing, China, July 30-31.
- Xiang, Yang, Wang, Xiaolong, Han, Wenying, and Hong, Qinghua. (2015). Chinese Grammatical Error Diagnosis Using Ensemble Learning. Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 99–104.

- Yeh, Jui-Feng, Yeh, Chan-Kun, Yu, Kai-Hsiang, Li Ya-Ting, and Tsai, Wan-Ling. (2015). Condition Random Fields-based Grammatical Error Detection for Chinese as Second Language. Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 105–110.
- Yu, Liang-Chih, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA 2014). 42-47.