CYUT-III Team Chinese Grammatical Error Diagnosis System Report in NLPTEA-2018 CGED Shared Task

Shih-Hung Wu*, Jun-Wei Wang
Chaoyang University of Technology,
Taichung, Taiwan, R.O.C
*Contact author: shwu@cyut.edu.tw

Liang-Pu Chen, Ping-Che Yang
Institute for Information Industry, Taipei, Taiwan, ROC.
{eit, maciaclark}@iii.org.tw

Abstract

This paper reports how we build a Chinese Grammatical Error Diagnosis system in the NLPTEA-2018 CGED shared task. In 2018, we sent three runs with three different approaches. The first one is a pattern-based approach by frequent error pattern matching. The second one is a sequential labelling approach by conditional random fields (CRF). The third one is a rewriting approach by sequence to sequence (seq2seq) model. The three approaches have different properties that aim to optimize different performance metrics and the formal run results show the differences as we expected.

1. Introduction

Learning Chinese as foreign language is getting popular. However, it is very hard for a foreign learner to write a correct Chinese sentence. We believe that a computer system that can diagnose the grammatical errors will help the learners to learn Chinese fast.

Since 2014, the NLP-TEA workshop provides a Chinese Grammar Error Detection (CGED) shared task to promote the research on diagnosis. The organizer provides learners' corpus tagged with error labels. There are four types of errors in the leaners' sentences: Redundant, Selection, Disorder, and Missing. The research goal is to build a system that can detect the errors, identify the type of the error, and point out the position of the error in the sentence (Yu et al., 2014). This year, the CGED added a new requirement: for errors of missing words and word selection, systems are required to recommend at most 3 corrections. If one of the corrections of these instances is identical with gold standard, the instances will be regarded as correct cases.

In 2018, we sent three formal runs in three different approaches. The first two are based on previous works, the first one is a pattern-based approach by frequent error pattern matching and language model scoring; the second one is a sequential labelling approach by conditional random fields (CRF), which performs well in year 2015 and 2016. The third one is a new approach, called rewriting approach by sequence to sequence (seq2seq) model. In the following sections, we will introduce the three approaches, discuss the formal run results, and give conclusion and future works.

2. Previous Works

2.1 Pattern-Based Approach

The pattern matching approach is an old approach, which has been used in many previous works (Wu et al., 2010; Chen et al., 2011). The pattern contains frequent error terms, in which a character is replace by a similar one. This is based on an assumption that students often make mistake among similar characters (Liu et al., 2009). The advantage of pattern matching is stable, the many drawback is it cost a lot to collect the patterns.

The system is based on the previous work, the error pattern from a native student essay corpus in traditional Chinese. Before testing

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 199–202
Melbourne, Australia, July 19, 2018.

the system, the test data is transformed into traditional by MS-Word 2010.

2.2 Sequential Labelling Approach

The second one is a sequential labelling approach by conditional random fields (CRF), which performs well in CGED 2015 and 2016.(Chen et al., 2015; Chen et al., 2016b)

The sequential labelling approach is based on the conditional random field (CRF) model (Lafferty, 2001). CRF has been used in many NLP applications, such as named entity recognition, word segmentation, information extraction, and parsing. To apply it to a new task, it requires a specific feature set and labeled training data. The CRF model is regarded as a sequential labeling tagger. Given a sequence *X*, the CRF can generate the corresponding label sequence *Y*, based on the trained model. Each label *Y* is taken from a specific tag set, which needs to be defined in each task. How to define and interpret the label is a task-depended work for the developers.

Mathematically, the model can be defined as:

$$P(Y|X) = \frac{1}{Z(X)} \exp(\sum_k \lambda_k f_k) \quad (1)$$

where *Z(X)* is the normalization factor, f_k is a set of features, λ_k is the corresponding weight which will be learned in the training process. In the CGED task, *X* is the input sentence, and *Y* is the corresponding error type label. We define the tag set as: {O, R, M, S, D}, corresponding to no error, redundant, missing, selection, and disorder respectively. Table 1: A sample of the CRF sequential labeling dataset shows a sample of our working file. The first column is the input sentence *X*, and the third column is the labeled tag sequence *Y*. Note that the second column is the Part-of-speech (POS) of the word in the first column. The combination of words and the POSs will be the features in our system. The POS set used in our system is a simplified POS set provided by CKIP[1].

Term	POS	Tag
可是	C	O
有	Vt	O
一點	DET	O
冷	Vi	O
了	T	R
你	N	O
的	T	R
過年	Vi	O
呢	T	O

Table 1: A sample of the CRF sequential labeling dataset

Since the system is based on the previous work, the training set is the 2014, 2015, and 2016 CGED training dataset in traditional Chinese. The test data is also in transformed into traditional by MS-Word 2010.

3. Rewriting Approach

This year, we propose a new approach, called rewriting approach. Given a sentence with grammar errors, a system can rewrite it and output a sentence without grammar error. This idea is inspired from the RNN encoder-decoder models, which have been used in many deep learning researches. In such models, with the help of a large training set, a sequence can be transformed into another corresponding sequence. Amount them Sequence-to-sequence (seq2seq) models (Sutskever et al., 2014, Cho et al., 2014) have been applied successfully to a variety of NLP tasks such as machine translation, speech recognition, text summarization and conversation generation (Wu et al., 2017). In this task, we also adopt the seq2seq model as it is in Neural Machine Translation (NMT) which was the very first testbed for seq2seq model.

3.1 Seq2seq Model

Our rewrite approach system is built on TensorFlow Sequence to sequence (Seq2Seq) model[2] with the long-short-term-memory (LSTM). The training set is the 2017 and 2018 CGED training dataset.

[1] http://ckipsvr.iis.sinica.edu.tw/

[2] https://www.tensorflow.org/tutorials/seq2seq

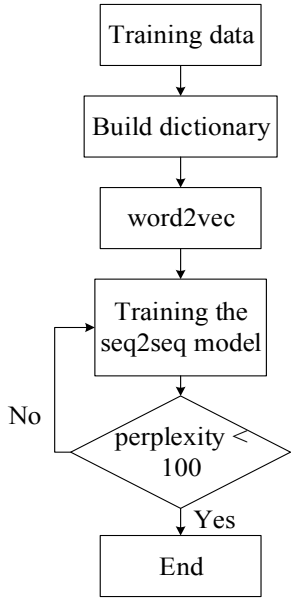

Figure 1. Training flowchart

3.2 Training the Seq2seq model

Figure 1 shows the training flowchart of our system. The first step is collecting all the vocabulary in the training corpus to build a dictionary. Then uses the word2vec model (Mikolov et al., 2013) to find the vector representation of each word. The sentences written by the students and the corresponding correct version sentences are used to train the seq2seq model. Since we do not have a validation set to find a better early stop point. The termination criterion of training is an empirical value, perplexity equal 100.

3.3 Preprocessing

The sentences are segmented by Jieba[3] word segmentation toolkit. The size of the vocabulary set is 5,424. The vocabulary is not very large, comparing to other the corpus used in other NLP tasks.

3.4 Post-processing

After the input is rewritten by the system, then the system will compare the rewritten sentence to the input sentence. We assume the rewritten one is the correct one and report the differences as grammar errors.

4. Experiment

4.1 Metrics

In the formal run, accuracy, precision, recall, and F-score are reported in three different levels. False positive rate is reported for the detection levels.

4.2 Formal Run result

The performance of our systems is shown in the following tables comparing to the average of all 32 formal runs in 2018. Table 2 shows the false positive rate; the only index that should be as low as possible. As we expected, the run1 pattern based approach gives the lowest FPR in all 32 runs.

Table 3 shows the performance evaluation

Submission	False Positive Rate
run1	**0.050**
run2	0.178
run3	1.000
Average of all 32 runs	0.467

Table 2: The false positive rate.

in detection level. At this level, the run2 sequential labelling approach perform well in both accuracy and precision. The recall is also improved from the performance in 2016 (Chen et al., 2016a). The performance of rewriting approach gives highest recall and

	Detection Level			
	Accuracy	**Precision**	**Recall**	**F1**
run1	0.468	0.695	0.090	0.159
run2	**0.602**	**0.754**	0.428	0.546
run3	0.473	0.581	**0.845**	**0.688**
Average of 32 runs	0.587	0.667	0.635	0.613

Table 3: Performance evaluation in Detection Level

high F1, however, poor accuracy and precision. This is also as we expected, since

[3] https://github.com/fxsjy/jieba

the training corpus is too small and the vocabulary size is also too small.

5. Conclusion and Future Works

This paper reports our approach to the NLP-TEA-5 CGED Shared Task evaluation. By comparing three different approaches, we find that the systems can be tuned to optimize different performance metrics.

Our system presents the best false positive rate in detection level by pattern matching approach and high accuracy, precision by sequential labelling approach and high recall and F1 by rewriting approach.

Due to the limitation of time and resource, our system is not tested under different experimental settings. In the future, we will use a larger corpus to train a better rewriting system to improve the performance on error diagnosis.

6. Acknowledgments

This study is conducted under the "III System-of-systems driven emerging service business development Project" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

Reference

Po-Lin Chen; Wu Shih-Hung; Liang-Pu Chen; Ping-Che Yang, (2016b) CYUT-III System at Chinese Grammatical Error Diagnosis Task, in Proceedings of The 3rd Workshop on Natural Language Processing Techniques for Educational Applications, Osaka, Dec 12, 2016.

Po-Lin Chen, Shih-Hung Wu, Liang-Pu Chen and Ping-Che Yang, (2016a) Improving the Selection Error Recognition in a Chinese Grammar Error Detection System, in Proceedings of the IEEE 17th International Conference on Information Reuse and Integration, July 28-30, 2016, Pittsburgh, PA, USA.

Po-Lin Chen; Wu Shih-Hung; Liang-Pu Chen; Ping-Che Yang; Ren-Dar Yang, (2015), Chinese Grammatical Error Diagnosis by Conditional Random Fields, in Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 7–14, Beijing, China, July 31.

Yong-Zhi Chen, Shih-Hung Wu, Ping-che Yang, Tsun Ku, and Gwo-Dong Chen (2011), Improve the detection of improperly used Chinese characters in students' essays with error model" Int. J. Cont. Engineering Education and Life-Long Learning, Vol. 21, No. 1, pp.103-116, 2011.

Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang and Shih-Hung Wu, (2009), "Capturing Errors in Written Chinese Words", in Proceedings of Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, Aug. 3-5.

Cho, K. et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proc. Conference on Empirical Methods in Natural Language Processing 1724–1734. arXiv preprint arXiv:1406.1078.

Lafferty, A. McCallum, and F. Pereira. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Intl. Conf. on Machine Learning.

Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781.

Ilya Sutskever and Oriol Vinyals and Quoc V. Le, (2014), Sequence to Sequence Learning with Neural Networks, Proc. NIPS. http://arxiv.org/abs/1409.3215.

Shih-Hung Wu, Yong-Zhi Chen, Ping-che Yang, Tsun Ku, Chao-Lin Liu, (2010), Reducing the False Alarm Rate of Chinese Character Error Detection and Correction, Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP), pages 54–61, Beijing, 28-29 Aug.

Shih-Hung Wu, Wen-Feng Shih, Che-Cheng Yu, Liang-Pu Chen, and Ping-Che Yang, (2017), CYUT-III Short Text Conversation System at NTCIR-13 STC-2 Task, in Proceedings of the 13rd NTCIR Conference on Evaluation of Information Access Technologies, Dec. 5-8, Tokyo Japan.

Yu, L.-C., Lee, L.-H., & Chang, L.-P. (2014). Overview of grammatical error diagnosis for learning Chinese as a foreign language. In Proceedings of the 1stWorkshop on Natural Language Processing Techniques for Educational Applications, 42-47.