A Hybrid Approach Combining Statistical Knowledge with Conditional Random Fields for Chinese Grammatical Error Detection

Yiyi Wang **Chilin Shih**
East Asian Languages and Cultures
University of Illinois at Urbana-Champaign
{ywang418, cls}@uiuc.edu

Abstract

This paper presents a method of combining Conditional Random Fields (CRFs) model with a post-processing layer using Google n-grams statistical information tailored to detect word selection and word order errors made by learners of Chinese as Foreign Language (CFL). We describe the architecture of the model and its performance in the shared task of the ACL 2018 Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA). This hybrid approach yields comparably high false positive rate (FPR = 0.1274) and precision (P_d= 0.7519; P_i= 0.6311), but low recall (R_d = 0.3035; R_i = 0.1696) in grammatical error detection and identification tasks. Additional statistical information and linguistic rules can be added to enhance the model performance in the future.

1 Introduction

Grammatical error detection is a growing area of research with general applications to grammar checking and Computer-Assisted Language Learning (CALL). NLPTEA shared task provides a platform for researchers to work on detecting the same types of grammatical errors, and evaluate the results on the same test set with predefined metrics(Yu et al., 2014; Lee et al., 2015, 2016; Rao et al., 2017) . Since NLPTEA 2014, the shared tasks focus on detecting and identifying four types of errors which are the most common grammatical mistakes made by CFL learners: word missing errors ("M"), word redundancy errors ("R"), word selection errors ("S"), and word ordering errors ("W"). The NLPTEA-2018 shared task focuses on identifying and correcting the above four types of errors made by CFL learners. The training data released by the task organizers contains 402 sentences written by Chinese language learners and corrected by native speakers of Chinese. The test data for the task consists of 3,548 sentences. The diagnose level evaluation metrics are based on three criteria: (1) detection-level: to distinguish grammatical and ungrammatical sentences; (2) identification-level: to identify error type; (3) position-level: to pin down error positions. Our model is designed to tackle the error detection task.

Most of the proposed methods for grammatical error detection employ supervised machine learning or deep learning approaches(Chen et al., 2016; Zheng et al., 2016; Chou et al., 2016) in recent years. Although neural networks model performs well for the complexity of the task in nature, CRFs still get steady application in the community. This paper proposes a integrated approach of combining CRFs, statistical information from Google n-grams and rule-based expert knowledge to detect the four types of errors. The method can yield high accuracy and precision, but low recall. To improve recall in the future, additional rules and statistical knowledge can be added to enhance model performance.

2 Data

In addition to the training data released by the task organizers, another data set containing 9,602 sentences with 23,518 types of grammatical errors employed in a similar shared task in NLPTEA 2016 is used in conjunction to train a CRFs model to detect all four types of grammatical errors. Table 1 is the distribution of the four types of errors in our training set.

Google Chinese Web 5-gram (Liu et al., 2010) is used to retrieve statistical information in the post-processing layer. The data is composed of around 883 millions of tokens generated from pub-

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 194–198
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

	NLPTEA 2018	NLPTEA 2016	Total
M	298	6,202	6,500
R	208	5,270	5,478
S	474	10,426	10,900
W	87	1,620	1,707

Table 1: Distribution of Errors in training set.

licly accessible web pages written in Chinese characters. Low frequency n-grams occurring less than 40 times are filtered out. However, some frequently occurring typos, ungrammatical forms, idiosyncratic usages, even texts written by language learners and/or written in other languages such as in Japanese Kanji are kept in the final published version of the data, making it challenging to identify the subtleties of non-native speakers' writings. For example, the word "坑生素(antibiotic)" occurs 200 times in the data,in which it contains one misused character "坑(pit)" that shares similarities in orthography with the correct usage "抗(anti)". So, when the form "坑生素" is used in CFL writing, it would pass the grammar checker based on Google n-gram due to its high frequency. Another example is "知情达理(understanding and reasonable)" with 10,495 occurrences in the data. This is a case of portmanteau combining two idioms "知书达礼(well-educated and courteous)" and "通情达理(show common sense)", in which the misused character "知(to know)" shares semantic component with the correct character "通(to go through)".

Although these entries are considered as noises in the Google n-grams collection, they provide exemplary language mis-usage information by CFL learners, and can bring in valuable insights about the typical grammatical errors made by CFL learners that we can use in grammatical error detection task. We will discuss how to use the information to identify word selection and word order error in Section 3.3.

3 Model Components

The model is designed to feed the sentences into a CRFs model to detect four types of grammatical errors, and pass the results to a post-processing layer to further identify word selection and word order errors based on unigram and bigrams information retrieved from Google Chinese n-grams. We describe the data preprocessing, feature sets selection of CRFs model, and post-processing step that modifies the CRFs output in the following sections.

3.1 Data Preprocessing

Since words are the basic element for many natural language processing tasks, and Chinese writing system by nature does not mark word boundaries, the first step of preprocessing is to segment the sentences into words. Stanford Word Segmenter is used to split the input sentences into sequences of words in terms of Peking University standard (Tseng et al., 2005) . Then the segmented sentences are fed into Stanford POS Tagger (Toutanova et al., 2003) to get parts of speech of each word. During the word segmentation and tagging processes, punctuations are treated as words, however, since they are not included in Google n-gram data, all the punctuations in the training set are removed to make the best use of available statistical information during the post-processing step. The sentences are presented as a three-column frame, with the first column as word, the second column as POS tagging, and the last one as error-detection output labels. Part of preprocessed training data is presented in Table 2.

Word	POS	Error
因此	AD	C
不仅	AD	C
靠	P	M
国家	NN	C
的	DEG	C
措施	NN	C
而且	AD	C
我们	PN	C
消费者	NN	C

Table 2: Example of preprocessed data.

3.2 Conditional Random Fields

CRFs (Lafferty et al., 2001) is a powerful model for predicting sequential labels with a wide range of applications in the NLP community, such as name entity recognition, POS tagging and parsing. The reason that CRFs is appropriate to model sequencing tasks is that it can take the contextual observations, usually a sequence of tokens as input and generates a sequence of labels as output, as in most of sequential labeling tasks.

The sequencing CRFs model, or linear chain CRFs, is well suited to the grammatical error detection task, as it can take the sentences as input sequences, and output the corresponding grammatical error labels. In our task, the output set is composed of five elements C, M, R, S, W, abbreviating for correct, missing, redundancy, selection and word ordering errors respectively.

CRFs provide a rich unconstrained feature set to represent data, and assigns a weight to each feature. Therefore, feature set construction can decide the expressive power of the model. We use 46 features in our model to represent the relationships between adjacent words, parts-of-speech, and their interaction in error prediction. CRF++ toolkit of Version 0.58 (**?**) is adopted in our model.

3.3 Post-Processing Layers

Two layers are added on top of the CRFs model to enhance performance by detecting grammatical errors based on the statistical information retrieved in Google Chinese n-grams. The first post-processing layer is applied to identify word selection error in terms of unigram information; the second layer is implemented to detect word-ordering error and word selection errors according to bigrams information.

3.3.1 Unigram Layer

The unigram layer applies to the words that are predicted as “C” in CRFs model to check the prediction accuracy by using unigram information; however, the words that are detected as errors will not be processed in this step. The post-precessing procedure of this step can be summarized as follows:

If a word is not a cardinal or ordinal number, the length of the word is not longer than two characters, and the occurrences of the word in Google unigram are less than 40,000 times, the original correct tag generated by CRFs is converted to a word selection error. The algorithm applied in this layer is shown Table 3 .

The rationale behind this design is that the frequencies of multisyllabic Chinese words decrease when their usages are unconventional. Therefore, when such expressions are found in CFL learner’s writing, there are reasonable grounds to believe that word selection errors have occurred.

Since the corpus cannot include all the numbers and proper nouns, the words with relatively low frequencies, such as a proper noun “栋杰(35,205)” and an ordinal number ”第三百三十九(39,982)” are likely to bee grammatical expressions. For this reason, parts-of-speech knowledge is integrated with the frequency information to better identify errors. The frequency threshold is decided by descriptive statistics of Google n-grams data. Although this setting improves the model recall in this task, the rationality of setting this cut-off will be discussed further in Section 4. In this step, if a word “灵恬(214)” or “快子(15,700)” is marked as “C” by CRFs model with a non “CD, OD or NR” POS tagging, the predicted tag is changed to “S”.

Algorithm 1: *Tag C is converted to Tag S based on unigram statistics*

```
if    (output = “C”    and
      POS !=“CD”, “OD” or “NR”    and
      wordLength <=2    and
      wordFrequency <=40,000):
      “C” is changed to “S”
```

Table 3: Unigram algorithm.

3.3.2 Bigrams Layer

This layer is used to further identify word selection and word order errors in terms of bigrams frequencies. If occurrences of bigrams are less than 1,000 times in the Google ngrams corpus, the range is detected as suspicious area that may contain grammatical errors. In this step, additional preprocessing is needed to chunk input sentences into bigrams with their corresponding frequencies in Google ngrams data. A preprocessed sentence as an example is shown in Table 4.

Since two words are contained in each suspicious area, the error type of individual word needs to be further decided. Unigram information is applied again to diagnose grammatical errors at the word level. The pseudo code used in this layer is presented Table 5.

If both of the words within the suspicious area have high word frequencies in the unigram data, such as “道(193,135,155)” and “吸烟(7,594,378)” in Row 3 of Table 4, the error may occur in the previous two words, if the previous bigrams also have low frequencies. In this case, both “道” and “吸烟” are correct words, however, the grammatical error occurs in the previous word “知不”. Simi-

Bigrams		Frequency
他们	知不	0
知不	道	0
道	吸烟	354
吸烟	对	153,530
对	未成年	98,312
未成年	年	461
年	的	91,329,920
的	影响	47,251,277
影响	会	324,577
会	造成	6,907,267
造成	的	20,711,377
的	各种	19,073,836
各种	害处	524

Table 4: Example of preprocessed data.

larly, this procedure can be applied to check following bigrams to decide the error type of individual word within a suspicious area.

Algorithm 2: *Tag C is converted to Tag W or Tag S based on bigrams and unigram*

```
if (word1Frequency >=40,000 and
    word2Frequency >=40,000 ):
    if previousBigramsFrequency>=1000:
        word1 is marked as "C"
        word2 is marked as "S"
    if postBigramsFrequency>=1000:
        word1 is marked as "S"
        word2 is marked as "C"
    else:
        word1 is marked as "C"
        word2 is marked as "C"
else:
    for $word_i$Frequency <40,000:
        if $word_i$Length>1:
            swap characters to new bigrams
            if newBigramsFreq >1000:
                $word_i$ is marked as "W"
            else:
                $word_i$ is marked as "S"
        else:
            if $word_i$POS == "CD":
                $word_i$ keeps the tag "C"
            else:
                $word_i$ is marked as "S"
```

Table 5: Bigrams algorithm.

If the frequency of at least one word within a suspicious area is less than 40,000, it is possible to assume that at least one grammatical error appears within this area. For example, the bigrams “他们 知不” in Row 1 of Table 3, since the word “知不” has zero occurrence in unigram, we can identify it is an error. Then we can swap the characters, get a new bigrams “他们 不知” and check the frequency of the new bigrams in the corpus. Since the frequency of “他们 不知” is 73,080, the word “知不” is marked as a word order error; otherwise, the low frequency individual word is marked as a selection error.

In this step, word order and selection errors are further detected in terms of both statistical information and linguistic knowledge. Table 6 shows an example of re-marked tags after passing this layer.

Word	CRFs Tag	Post-processed Tag
他们	C	C
知不	S	W
道	C	C
吸烟	C	C
对	C	C
未成年	C	C
年	R	R
的	R	R
影响	C	C
会	C	C
造成	C	C
的	C	C
各种	C	C
害处	C	S

Table 6: Example of post-processed tags.

4 Results and Discussions

The model yields high precision, but low recall in the shared task. The detailed evaluation results are shown in Table 7.

Since the post-processed layers are designed to detect word selection and word order errors only, considering the large amount of word missing and redundancy errors in the test data, it is expected that some false negative elements are failed to be identified in this model. In the future, more statistical information and linguistic rules can be added to reinforce the performance of this hybrid model.

	Precision	Recall	F1
Detection	0.7519	0.3035	0.4324
Identification	0.6311	0.1696	0.2673
Position	0.2385	0.0536	0.0875

Table 7: Test results of hybrid model.

The evaluation results of using CRFs alone and the hybrid model we proposed are compared in Table 8. By adding the post-processed layer, there is a trade-off between precision and recall. The decrease in precision is possibly caused by the increase of false positive errors, because words with frequencies lower than 40,000 are marked as selection errors in the post-processed layer. Some words, such as “幽默性(19,928)” and “梧桐花(37,707)”, even though with low frequencies, are grammatical expression in Chinese; however, they are identified as errors in the model by chance.

	Precision	Recall	F1
CRFs	0.8804	0.1444	0.2481
Hybrid	0.7519	0.3035	0.4324

Table 8: Comparison of CRFs model and hybrid model.

For the parameters setting in the post-processed layer, our model use 40,000 as the threshold for unigram, and 1,000 for bigrams. These two numbers are reached by observing the descriptive statistics of the data. Detailed corpus studies about the data distribution in Google n-grams can facilitate the parameter setting and in turn lead to better model performance in the future.

Since the post-layer is independent of the base model, it can be easily applied on top of other models, such as statistical, rule-based or hybrid models, to further promote the base model performance.

References

Po-Lin Chen, Shih-Hung Wu, Liang-Pu Chen, et al. 2016. Cyut-iii system at chinese grammatical error diagnosis task. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 63–72.

Wei-Chieh Chou, Chin-Kui Lin, Yuan-Fu Liao, and Yih-Ru Wang. 2016. Word order sensitive embedding features/conditional random field-based chinese grammatical error detection. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 73–81.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lung-Hao Lee, Gaoqi Rao, Liang-Chih Yu, Endong Xun, Baolin Zhang, and Li-Ping Chang. 2016. Overview of nlp-tea 2016 shared task for chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 40–48.

Lung-Hao Lee, 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 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2015)*, pages 1–6.

Fang Liu, Meng Yang, and Dekang Lin. 2010. Chinese web 5-gram version 1 ldc2010t06. CD-ROMs.

Gaoqi Rao, Baolin Zhang, XUN Endong, and Lung-Hao Lee. 2017. Ijcnlp-2017 task 1: Chinese grammatical error diagnosis. *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 1–8.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

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 1st Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–47.

Bo Zheng, Wanxiang Che, Jiang Guo, and Ting Liu. 2016. Chinese grammatical error diagnosis with long short-term memory networks. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 49–56.