WORD ORDER CORRECTION FOR LANGUAGE TRANSFER USING RELATIVE POSITION LANGUAGE MODELING

Chao-Hong Liu, Chung-Hsien Wu, and Matthew Harris

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan
chl@csie.ncku.edu.tw, *chwu@csie.ncku.edu.tw, matt@csie.ncku.edu.tw

ABSTRACT

Sentence correction has been an important and emerging issue in computer-assisted language learning. However, existing techniques based on grammar rules or statistical machine translation are still not robust enough to tackle the common incorrect word order errors in sentences produced by second language learners of Chinese. In this paper, a novel relative position language model is proposed to address this problem, for which a corpus of erroneous English-Chinese language transfer sentences along with their corrected counterparts is created and manually judged by human annotators. Experimental results show that compared to a scoring approach based on an n -gram language model and a phrase-based machine translation system, the performance in terms of BLEU scores of the proposed approach achieved improvements of 20.3% and 26.5% for the correction of word order errors resulting from language transfer, respectively.

Index Terms— Chinese as a Second Language, Language Transfer, Relative Position Language Modeling

1. INTRODUCTION

Automatic error detection and correction of sentences produced by second language learners has been an emerging field in computer-assisted language learning (CALL) [1-4]. Although error correction and detection of erroneous sentences are two related research topics that can work complementarily, they are considered as two distinct problems in contemporary approaches.

Specific error types have been identified for detection and correction by different researchers, mainly for the English language. In the ALEK system (Assessing LEXical Knowledge) developed by [2], detection of a sentence with errors is addressed using 20 target words extracted from the Test of English as a Foreign Language (TOEFL). In [1], the errors for correction, according to the Japanese Learner's English corpus [3], focus on those involving specific parts-of-speech such as articles and number agreement. In [5], the countability errors encountered in English as a Second

Language (ESL) are considered for correction using the Chinese Learner English Corpus (CLEC).

There are several approaches to the problem of error correction, or grammar correction, for second language learners. Natural language generation (NLG) from meaning representations is one possible direction to deal with the sentence correction problem [1]. However, since it requires in-depth analysis on not only the semantics of sentences but also the semantics of ill-formed sentences, it is not currently feasible to yield the meaning representations for the purpose of error correction [6-8]. NLG from keywords could be another possible direction [9]. However, it does not seem applicable to the word order correction problem in which we are interested.

An intermediate method between the above two NLG approaches is presented in [1], which we refer to as the Stripping and Insertion of Limited Candidates (SILC) approach. This correction procedure is language-specific (to English) and does not address the word order correction problem. Furthermore, the Chinese counterparts of the considered insertion candidates of articles, prepositions, and auxiliaries do not comprise the major errors occurring in sentences of English-Chinese language transfer, due to the characteristics of the Chinese language [10].

Recently statistical machine translation (SMT) seems to be a promising means for error correction, where the problem of word order correction could be solved within the phrase-based framework [11], with a pilot study presented by [5]. It is important to note that SMT has the potential to solve any kind of errors occurring in sentences produced by second language learners. The error sentences and their corresponding valid sentences can be considered as the source and the target, respectively. As a preliminary evaluation, only errors associated with mass nouns are considered for correction in [5]. It is also important to note that current state-of-the-art phrase-based machine translation systems depend heavily on n -gram language model and there is little long-range lexical information (which we believe is useful for the word order correction problem) incorporated into the process of ranking translated (corrected) sentences.

In this paper, we present a novel approach using the proposed relative position language model to deal with word

order errors, which comprise about one third of the errors made by learners of Chinese as a Second Language. To the best of our knowledge, this is the first research dedicated to the automatic correction of word order errors for second language learning.

The organization of this paper is as follows. Section 2 presents the scope of errors considered in this paper and Section 3 gives an overview of the approach. Section 4 details the proposed relative position language model and its formalisms. Section 5 describes the experimental setup of an n -gram based and an SMT-based baseline systems and the proposed method, and gives the results and comparisons of these approaches. Finally, we present the conclusions and future directions of this research.

2. WORD ORDER CORRECTION

Acknowledging the significance of language transfer and its influence on second language learning, we observe from a small corpus of English-Chinese error sentences that most of the errors made by Chinese learners can be attributed to negative language transfer, i.e., mother tongue influence. Of those errors, in contrast to the errors commonly made by learners of English as a Second Language (ESL) such as the use of articles and prepositions, we found in the corpus that Chinese learners tend to produce sentences with word order, lexical choice, insertion, and deletion errors.

Since word order errors alone exist in about 45.3% of all the error sentences produced by Chinese learners, and there is no existing technique dedicated to resolve this kind of errors except a more generalized SMT approach, in this paper we only focus on the word order errors arising from language transfer. Fig. 1 shows an example sentence with two word order errors and their corresponding corrections.

The original English sentence:

I went to sleep at 12 last night

The Chinese sentence with word order errors from language transfer: 我睡覺十二點昨天晚上

Correction 1 (no errors): 我昨天晚上十二點睡覺

Correction 2 using SMT: 昨天晚上十二點，我睡覺

Fig. 1: An example Chinese word order error sentence arising from language transfer and its two automatic corrections; the first correction is produced using the proposed method and is a correct sentence, and the second is made by an SMT system and is not well corrected in this example.

3. SYSTEM OVERVIEW

Most of the research in CALL considers only the detection of grammatical errors [2-4, 12]. Of those approaches dedicated to error correction, the error types involved are usually specific to the grammar of the English language [1, 5]. Furthermore, the detection and correction of errors are considered as two distinct problems in previous work [1, 5].

Therefore existing procedures employed for error correction do not detect if a sentence contains an error in the first step; all sentences are regarded as erroneous for correction. We, instead, incorporate an error detection module in the proposed error correction procedure, as shown in Fig. 2.

The main difference between the proposed procedure and those in [1] and [5] is 1) an error detection module is incorporated in our framework, and 2) the correction of word order errors in the proposed method involves long-range re-structuring of words in sentences, while the entities considered to be corrected in previous research are individual words.

Fig. 2: System overview of error sentence correction with an error detection module for language transfer.

In the error detection phase, a classifier is trained to detect if a sentence contains a word order error. Then, in the error correction phase, similar to SMT, the noisy channel model is also used in the present work to yield the corrected sentences [5]. Given a sentence E with one or more word order errors, we seek to return the most probable corresponding corrected sentence with:

$$\begin{aligned}\hat{C} &= \arg \max_C P(C|E) \\ &= \arg \max_C P(E|C) \cdot P(C)\end{aligned}\quad (1)$$

where C is a corrected sentence candidate, $P(E|C)$ is the reordering model and $P(C)$ is the probability of C using a language model derived from a large corpus of correct sentences.

The reordering model $P(E|C)$ is used to estimate the transformation probability of a reordered sentence with respect to the input error sentence E . In the present work we use the Probabilistic Context Free Grammar (PCFG) probability of C as the structural transformation probability for reordering modeling.

For the purpose of word order correction for language transfer, we believe that the language model plays a key role in determining the final corrected sentence. Since n -gram language model is unsuitable for capturing long-range lexical relationship in a sentence, it is inappropriate for the word order correction problem. Therefore, in the present work we propose a novel language modeling for $P(C)$ as described in Section 4.

4. RELATIVE POSITION MODELING

In designing an appropriate language modeling technique for word order correction, the relative position relationships

and the long-range lexical information between the constituent words in a sentence are equally important. With these guidelines, the *positional score* of a word w_a with respect to another word w_b across a corpus is first given by:

$$r(w_a | w_b) = \frac{\text{count}_{RP}(w_a, w_b)}{\text{count}(w_a, w_b)} \quad (2)$$

where $\text{count}(w_a, w_b)$ is the number of sentences in a corpus in which w_a and w_b co-occur, and $\text{count}_{RP}(w_a, w_b)$ is the number of sentences in the corpus where w_a and w_b co-occur and w_b occurs before w_a . It should be noted that despite appearances, the term $r(w_a | w_b)$ is not a probability measure, even though in fact $r(w_a | w_b) + r(w_b | w_a) \geq 1$.

By relative position, what is of interest is to capture the likelihood of a word w_b appearing to the left of w_a , without considering all the words (if any) in between, by which the long-range lexical information can be incorporated.

Given the positional score of each word, the relative position score of $C = w_1 w_2 \dots w_n$ is then defined as:

$$R(C) = R(w_1 w_2 \dots w_n) = \frac{\sum_{i=2}^n \left[\sum_{j=1}^{i-1} r(w_i, w_j) \right] / (i-1)}{n-1} \quad (3)$$

where n is the number of words in C .

To capture the distinguishing property of the sentences that have language transfer errors corrected and those that have not, a weighted relative position score is used as the language model $P(C)$ in this approach, and is defined as:

$$P(C) = \frac{R_{REF}(C)}{R_{ERR}(C)} \cdot R(C) \quad (4)$$

where $R_{REF}(C)$ and $R_{ERR}(C)$ are both relative position scores; the difference being that $R_{REF}(C)$ and $R_{ERR}(C)$ are estimated respectively from the reference sentences and error sentences of a parallel error corpus, while $R(C)$ is estimated from a large corpus with correct sentences.

5. SYSTEM EVALUATION

To evaluate the word order correction performance of the proposed method (denoted as the P-System), we used two baseline systems for comparison. The first one used an n -gram language model as opposed to the proposed method, and is denoted as B1-System. The baseline system for the Shared Task of the ACL 2007 Workshop on Statistical Machine Translation (www.statmt.org/wmt07/baseline.html) was employed as our second baseline system, and is denoted as B2-System. The main components of this SMT system are 1) n -gram language modeling using SRILM, 2) word and phrase alignment using GIZA, mkcls, and GIZA++, and 3) decoding using Moses.

5.1. Training and Test Datasets

The parallel text used to train and test these systems was prepared by combining three parallel corpora: 1) Chinese English News Magazine Parallel Text (LDC2005T10, denoted as Sinorama), 2) the ISI Chinese-English Automatically Extracted Parallel Corpus (LDC2007T09, denoted as ISI), in which only sentences of length less than 25 are used, and 3) pairs of Chinese-English example sentences (constructed for pedagogic purposes) extracted from the Dr.eye electronic dictionary (www.dreye.com.tw), denoted as DrEye).

Due to the lack of word order error sentence pairs, we created the English-ordered Chinese sentences using the word and phrase alignment information provided by GIZA++ from both the ISI and DrEye corpora, following a similar setup used in [5]; Sinorama was only used for SMT training since the sentence are more complex and the English-ordered Chinese sentences produced may not reflect the real word order errors made by second language learners. The DrEye corpus is divided into three subsets; each of the first two subsets contains 2,000 sentences (one for development testing for the B2-System, and the other for system evaluation/testing for all the three systems) and the remaining sentences together with Sinorama and ISI comprise the training set for B2-System. The fundamental information of the resulting corpus is shown in Table 1.

Table 1: The parallel corpus used for training and test.

	# sentence pairs	# words (avg./sen.)
Sinorama	77,859	1,036,188 (13)
ISI	33,681	401,685 (12)
DrEye	67,964	597,984 (09)
Total	179,504	2,035,857 (11)

For the training of language models for all these systems (5-grams using Chen and Goodman's modified Kneser-Ney discounting for the baseline systems, and positional scoring using the proposed Equation 2 for the P-System), a subset of about 10% of LDC2005T14 Chinese Gigaword (using file names ending with "0.tag") was used to reduce the required training time and memory space.

The probabilities of the grammar rules needed for calculating the PCFG for $P(E|C)$ were estimated from the automatic parses of the Sinorama corpus using the Stanford Parser (nlp.stanford.edu/software/lex-parser.shtml).

For the detection of word order errors, we used SVM^{light} (<http://svmlight.joachims.org/>) to train an SVM classifier using the n -gram, PCFG and the proposed relative position scores as the features of a sentence. A corpus containing 2,310 human-annotated error sentences (as positive samples) and 3,000 non-error sentences (as negative samples) were prepared for the training of the SVM.

5.2. Experimental Results

For error detection, we used the 10-fold cross-validation method to evaluate the performance of detecting if a sentence contains an error. The overall detection accuracy achieved by the trained SVM classifier was 96.7%. Sentences classified as containing no error do not undergo the error correction procedure as shown in Fig. 2.

The results of word order error correction using n -grams as the language model and the proposed method with relative position modeling are shown in Table 2, where the BLEU score was used as the evaluation metric [13]. For an input error sentence, each system outputs 200 corrected sentence candidates which are then re-ranked using the respective methods (B1-System using n -gram and P-System using the proposed method). The Top-N header shown in Table 2 denotes that the score is estimated using the maximum score of the Top-N scores of the re-ranked sentences. The results show that the proposed method outperforms the n -gram based approach in all the Top-N evaluations, and gives a near 0.10 improvement for the Top-1 and more than 0.05 for the Top-5 evaluation.

Table 2: Comparison of word order correction performance using n -gram-based B1-System and the proposed method.

BLEU	Top-1	Top-5	Top-10	Top-20
B1-System	0.468	0.707	0.796	0.873
P-System	0.563	0.761	0.833	0.889

The trained SMT-based B2-System reported 0.445 BLEU score, which is similar to the result of B1-System because they both are based on n -gram language model. The results show that by considering the relative position information for word order correction, the performance of the proposed P-System in terms of BLEU score can be improved by 20.3% and 26.5% compared to n -gram and SMT-based baseline systems, respectively.

6. CONCLUSIONS

We have described a framework to detect and correct sentences with word order errors for second language learners. The contributions of this work are:

1. It is the first NLP research trying to fix word order errors for language transfer to the best of our knowledge.
2. It is also the first work dedicated to fixing grammatical errors for the Chinese language.
3. A novel relative position language modeling technique is proposed to correct word order errors arising from language transfer to Chinese.

The experimental results show that the proposed method has a 26.5% improvement in terms of BLEU score compared to a state-of-the-art machine translation system trained on the same dataset.

Future directions of this work include 1) a more detailed analysis on the theoretical basis of the proposed relative

position language model with n -gram and other newly proposed language models, and 2) to extend the detection and correction procedure to other common error types made by Chinese as a Second Language learners, such as lexical choice, deletion, and insertion errors.

7. REFERENCES

- [1] J. Lee and S. Seneff, "Automatic Grammar Correction for Second-Language Learners," *Ninth International Conference on Spoken Language Processing*, 2006.
- [2] M. Chodorow and C. Leacock, "An unsupervised method for detecting grammatical errors," *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pp. 140-147, 2000.
- [3] E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara, "Automatic error detection in the Japanese learners' English spoken data," *Proc. ACL*, 2003.
- [4] E. M. Bender, D. Flickinger, S. Oepen, A. Walsh, and T. Baldwin, "Arboretum: Using a Precision Grammar for Grammar Checking in CALL," *InSTIL/ICALL Symposium*, 2004.
- [5] C. Brockett, W. B. Dolan, and M. Gamon, "Correcting ESL errors using phrasal SMT techniques," *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 249-256, 2006.
- [6] K. Knight and V. Hatzivassiloglou, "Two-level, many-paths generation," *Proc. ACL*, vol. 95, pp. 252-260, 1995.
- [7] I. Langkilde and K. Knight, "Generation that exploits corpus-based statistical knowledge," *Proceedings of the 36th annual meeting on Association for Computational Linguistics, August*, pp. 10-14, 1998.
- [8] A. Ratnaparkhi, "Trainable methods for surface natural language generation," *ACM International Conference Proceeding Series*, vol. 4, pp. 194-201, 2000.
- [9] K. Uchimoto, H. Isahara, and S. Sekine, "Text generation from keywords," *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1-7, 2002.
- [10] Y. Wang and R. Garigliano, "An Intelligent Language Tutoring System for Handling Errors Caused by Transfer," *Intelligent Tutoring Systems: Second International Conference, ITS'92, Montreal, Canada, June 10-12, 1992, Proceedings*, 1992.
- [11] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 48-54, 2003.
- [12] C. Leacock and M. Chodorow, "Automated Grammatical Error Detection," *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pp. 195-207, 2003.
- [13] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, 2001.