



Prosodic Word-Based Error Correction in Speech Recognition Using Prosodic Word Expansion and Contextual Information

Chao-Hong Liu and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Taiwan

ch1@csie.ncku.edu.tw, chunghsienwu@gmail.com

Abstract

In this study, considering the effect of phrase grouping in spontaneous speech, prosodic words, instead of lexical words, are adopted as the units for error correction of speech recognition results. The prosodic words and the corresponding mis-recognized word fragments are obtained from a speech database to construct a mis-recognized word fragment table for the extracted prosodic words. For each word fragment in a recognized word sequence, the potential prosodic words which are likely to be misrecognized as input word fragments are retrieved from the table for prosodic word candidate expansion. The prosodic word-based contextual information, considering substitution and concatenation scores, is then employed into a probabilistic model to find the best word fragment sequence as the corrected output. Experimental results show that the proposed method achieved a 0.32 F_1 score, with improvements of 0.18 and 0.10 compared to the SMT-based and lexical word-based approaches, respectively.

Index Terms: speech error handling, prosodic words, error correction, contextual information

1. Introduction

Due to the large variability in spontaneous speech, speech recognition process constitutes the major source of errors in most spoken dialogue systems. Accordingly, error correction of speech recognition results is highly desirable for the application to spoken dialogue systems [1, 2]. Error correction of automatic speech recognition (ASR) results is the technique incorporating linguistic or high-level features with which the computational complexity is too heavy to be considered in the recognition phase. The current state-of-the-art ASR systems for spontaneous speech recognition generally achieve recognition rates from 70% to 80% and usually consider mainly the acoustic and *bi*-gram information. In error correction of speech recognition results, lexical word-based candidate expansion followed by linguistic or semantic processing is generally used. However, with the investigation on the speech recognition results, the recognition errors are highly affected by their neighboring words.

In spontaneous speech, speech pronunciation is significantly influenced by phrase grouping. The constitution of phrases under grouping denotes the possible cognitive units of spontaneous speech in which articulation should interact with articulatory apparatus and breathing. Prosodic words are typically characterized by word stress and segmental word-level rules and therefore can be regarded as the basic units for the constitution of the spontaneous speech.

Since the grouping of prosodic phrases/words indicate the change of rhythm in a sentence and breath groups are often marked at prosodic boundaries, word fragments in a prosodic word tend to be mis-recognized together. Accordingly, in this

research, prosodic words are selected as the basic unit for error correction of speech recognition results. Furthermore, the prosodic word-based contextual information is employed into the linguistic level correction.

Several methods including transformation-based learning and memory-based learning were employed to correct ASR errors for spoken dialogue systems [1]. These techniques regarded only words and word sequence as correction units and conceptually there is no structural information considered in the correction process.

Recently structural information has been incorporated for error correction. In [2], grammar rules are taken as "Correction Grammar" in error handling. Grammar rules are also used in the field of computer-assisted language learning (CALL) to detect and correct errors made by second-language learners [3]. In [4], rich linguistic information is being used to form semantic, syntactic, lexical and contextual patterns as correction units. It should be noted that all these four kinds of patterns are shallow and are carefully devised for several sentence types such as "product order" and "telephone number" they are tested on.

In previous approaches, the techniques proposed for ASR error correction were usually applied to spoken dialogue systems in order to improve dialogue success rate, rather than simply correct the ASR errors [1, 2, 4]. This particular application has shaped the direction of ASR error correction in these systems. As a result, the information about the dialogue corpus being tested has served as prior knowledge on how the errors can be corrected in an ASR output sentence.

Recently it was shown that statistical machine translation (SMT) technique can be applied to correct errors by preparing appropriate parallel corpus of erroneous and correct texts [5-7]. SMT-based method also has the potential to correct most errors in ASR outputs, although it is not popularly employed to this application to date.

Structural information such as syntactic trees was employed both as proposing correction candidates and as means of fluency measure for the correction of errors in sentences written by second-language learner [6, 7]. This information is well suited for CALL systems because the sentences do not contain errors resulted from ASR. Second Language learners do not make mistakes that will completely change the overall sentence structure. Therefore, the syntactic tree structures are mostly the same in terms of their sub-tree structures.

However, the syntactic structures of a correct sentence and its ASR output sentence can be very different in many cases. To record the two syntactic sub-trees as a correction template is a viable solution, however, it will be too specific such that it is almost useless to correct the ASR errors in practice.

In this study, since the grouping of prosodic phrases/words is closer to the breathing groups and the rhythm of the sentence, the word fragments in a prosodic word is tentatively

to be mis-recognized together. For a spoken dialogue system, the dialogue corpus are collected and used to extract the prosodic words and the corresponding mis-recognized word fragments using a speech recognizer. A mis-recognized word fragment table recording all prosodic words and the corresponding mis-recognized word fragments is constructed. In the correction process, for each word fragment in a recognized word sequence, the potential prosodic words which are likely to be mis-recognized as the input word fragment are retrieved from the mis-recognized word fragment table for prosodic word candidate expansion. The prosodic word-based contextual information is then employed into a probabilistic model to determine the best word fragment sequence as the corrected output.

The organization of this paper is as follows. Section 2 provides the modeling for speech error handling, incorporating the idea of using prosodic words. The definition of prosodic words is given in Section 2.1. Two scores for prosodic words, the substitution score and the concatenation score, and their combined correction score for the whole sentence are also defined in Section 2.2. In Section 3 the experimental setup, results and comparisons are described. Finally in Section 4 the contributions of this research are concluded.

2. Speech Recognition Error Correction

The proposed model for speech error correction starts from a given input speech S , and the goal of this study is to obtain the corresponding “corrected” text C . The overall model can be written as

$$\hat{C} = \arg \max_c P(C|S) \quad (1)$$

Considering ASR speech confidence and textual correction factors, this model can be split into two sub-models by introducing a hidden variable E , which is composed of all possible error text outputs from an ASR. Therefore the model is rewritten as

$$\hat{C} = \arg \max_c \sum_E P(C|S,E)P(E|S) \quad (2)$$

In this research, only the most probable ASR text output is considered. Since speech signal S has been converted into all possible error text outputs E from an ASR, the model can be approximated by

$$\hat{C} \cong \arg \max_c \left(\max_E P(C|E)P(E|S) \right) \quad (3)$$

where the value of the term $P(E|S)$ can be obtained from the recognition confidence score of S in deriving E in the ASR system.

2.1. Prosodic Word Extraction

Fig. 1 depicts an example of lexical (upper part) and prosodic structures (lower part) for a Mandarin sentence. Compared to lexical structure, prosodic structure represents more intonation information of speech. The grouping of prosodic phrases/words is closer to the breathing groups and the rhythm of the sentence. In Mandarin speech, the prosodic word and phrase breaks often occur in lexicon word boundaries, though the lexicon word boundaries may not always be prosodic word/phrase breaks. Prosodic Structure is constructed by first predicting the prosodic phrase breaks and prosodic word breaks from the input sentence using the algorithm proposed in [8]. In the training corpus, each boundary between two

consecutive syllables is labeled as either breaks or non-breaks of prosodic phrases/words. A supervised classification and regression tree (CART) for prosodic word prediction is then constructed. First, all the syllable boundaries are used for PW break prediction. Finally, all the syllable boundaries are categorized into PW breaks or non-breaks. The contextual features for each boundary between two consecutive syllables are extracted in a window size of five syllables (three before and two after the boundary). The contextual features are extracted by a text analyzer which integrates the word segmentation and phonological analysis modules proposed by Wu and Chen [9].

To maximize the purity in child nodes at each splitting, the CART is trained by maximizing the reduced entropy ratio (RER), which is calculated as:

$$RER = \frac{E_{parent} - \sum_i (n_i/n_{parent}) E_i}{E_{parent}} \quad (4)$$

where n_{parent} and n_i are the numbers of the training data in the parent node and the i^{th} child node, respectively, and E_{parent} and E_i are the corresponding entropies. In each node, the entropy E is estimated as:

$$E = - \sum_{q=1}^Q p_q \times \log(p_q) \quad (5)$$

where p_q is the probability for observing class q in the node. Q denotes the total number of classes.

Based on the above procedure, the prosodic words for a dialogue system can be extracted from the collected corpus.

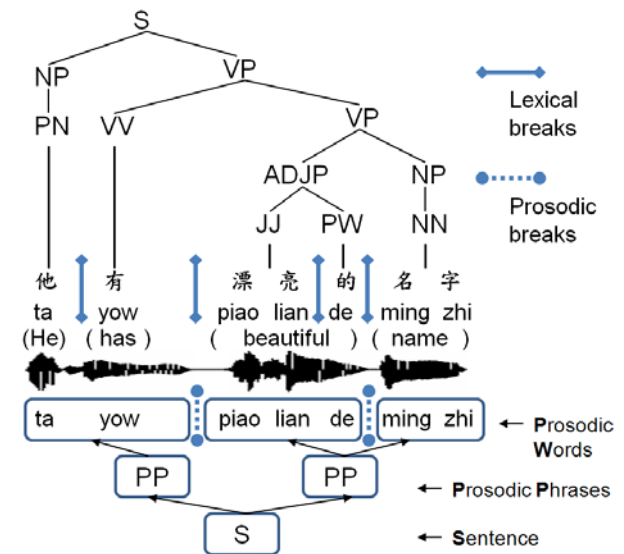


Figure 1: Prosodic breaks vs. Lexical breaks.

2.2. Mis-Recognized Word Fragment Table based on Prosodic Words

Based on the extracted prosodic words for a spoken dialogue system, the prosodic words which are likely to be mis-recognized as a word fragment are collected to construct a Mis-Recognized Word Fragment (MWF) Table using the collected spoken dialogue corpus. For example, prosodic word “漂亮的 (beautiful)” (piao lian de) is likely to be mis-recognized as the word fragments “聊聊個” (liao liao ge) and “六人的” (liou ren de) along with the occurrence frequencies as shown in Fig. 2. To construct the MWF table, similar to the

3. Experiments

3.1. Experimental Setup

For the evaluation of the proposed approach, one spoken dialog corpus for the travel information task was collected. This corpus collected the speech utterances from six male and two female subjects. The utterance collection system was basically a spoken dialogue system (SDS) that could interact with the speakers by asking/answering questions and recording the speakers' utterances. Totally, 144 dialogues consisting of 1586 sentences were collected in the travel information domain. In addition, the TH-CoSS corpus [10] was adopted to construct the CART for prosodic word extraction. An HTK3-based ASR system was employed to obtain the recognition results and their corresponding confidence scores. An SMT-based baseline system for the Shared Task of the Fourth Workshop on Statistical Machine Translation (www.statmt.org/wmt09/) was also constructed for comparison purposes.

3.2. Results and Discussions

Table 1 lists the comparison results of the proposed method and the SMT-based baseline system in terms of BLEU score and F_1 measure. The value listed within the parenthesis is the average BLEU score of the speech recognized texts as compared to their corresponding correct texts over the corpus, which can be served as a reference value for the BLEU scores calculated from the derived texts of the two approaches. If the BLEU score decreases, it essentially indicates that the correction is detrimental to the recognized texts as a whole.

We can see from Table 1 that corrected texts yielded from the SMT-based method actually have a lower BLEU score than the speech recognized texts (0.437 vs. 0.491). This indicates that as a whole, the corrected texts derived by SMT are in fact more "incorrect" than the ASR recognized texts. However, it was found that some corrections are actually of very high quality, although the BLEU score is low considering all the corrections.

As a comparison, the proposed method using prosodic words obtained a 0.497 BLEU score, which not only outperformed the SMT-based approach by 0.06 but is also larger than that of the ASR recognized texts by 0.006. Although the improvement from 0.491 to 0.497 is only small, it still represents that the proposed method is a viable approach to correcting the errors, rather than providing corrections which are more incorrect than the ASR recognized texts.

Table 1. Comparison of error correction results using SMT-based, Lexical word-based and the proposed methods

	SMT	Lexical-Word	Proposed
BLEU (0.491)	0.437	0.472	0.497
Precision	0.146	0.224	0.335
Recall	0.149	0.234	0.315
F_1	0.147	0.229	0.325

The precision and recall rates are estimated in terms of the correction of prosodic words. It is clearer to see also from these three scores that the proposed method outperformed the SMT-based approach, with an improvement of 0.18 in F_1 measure. By observing the correction results of the SMT, we found that the SMT approach only improved a little because:

(1) the correction pairs (aligned translation pairs in SMT terminology) are not of very high quality because the parallel corpus is not sufficiently enough for the alignment training for SMT, and (2) the n -gram language model applied to score the corrected sentence tends to ensure their high fluency, whereas the best correction candidates may not have such characteristics. These results have shown that the SMT-based approach may not be a suitable solution to the correction of ASR results. We will be examining its performance if more training data can be used in our future work.

When compared to the lexical word-based method, the proposed method still presents an improvement of 0.096 in F_1 measure. By observing the resulting corrections we found that if the correction considers lexical words as the correction units, there were too many unnecessary word corrections being applied to the final correction of the sentence, thus lowered the precision and recall scores. This result has confirmed the effectiveness of using prosodic words as the units for error correction of speech recognition results.

4. Conclusions

In this research, an error correction model based on prosodic words was proposed. A mis-recognized word fragment table and prosodic word expansion were employed. The prosodic word-based contextual information, considering substitution and concatenation scores, is then employed into a probabilistic model to find the best word fragment sequence as the corrected output. Experimental results show that the proposed error correction method for ASR outputs achieved a 0.18 F_1 improvement compared to the SMT-based method for the dialogue system in the travel information domain.

5. References

- [1] G. Skantze, "Error Handling in Spoken Dialogue Systems," in *Computer Science and Communication Department of Speech, Music and Hearing*. Stockholm, Sweden: KTH, 2007.
- [2] H. Sagawa, T. Mitamura, and E. Nyberg, "Correction grammars for error handling in a speech dialog system," *HLT/NAACL, Boston*, 2004.
- [3] E. M. Bender, D. Flickinger, S. Oepen, A. Walsh, and T. Baldwin, "Arboretum: Using a Precision Grammar for Grammar Checking in CALL," in *Proceedings of InSTIL/ICALL Symposium on Computer Assisted Learning*, 2004.
- [4] R. López-Cózar and Z. Callejas, "ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information," *Speech Communication*, vol. 50, pp. 745-766, 2008.
- [5] C. Brockett, W. B. Dolan, and M. Gamon, "Correcting ESL Errors using Phrasal SMT Techniques," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 2006, pp. 249-256.
- [6] C.-H. Liu, C.-H. Wu, and M. Harris, "Word Order Correction for Language Transfer using Relative Position Language Modeling," in *Proceedings of ISCSLP-2008* Kunming, China, 2008.
- [7] C.-H. Wu, C.-H. Liu, M. Harris, and L.-C. Yu, "Sentence Correction Incorporating Relative Position and Parse Template Language Models," *IEEE Trans. on Audio, Speech and Language Processing*, to appear in 2010.
- [8] M. Chu and Y. Qian, "Locating boundaries for prosodic constituents in unrestricted Mandarin texts," *Computational Linguistics and Chinese Language Processing*, vol. 6, pp. 61-82, 2001.
- [9] C.-H. Wu and J.-H. Chen, "Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis," *Speech Communication*, vol. 35, pp. 219-237, 2001.
- [10] W. Zhu, W. Zhang, Q. Shi, F. Chen, H. Li, X. Ma, and L. Shen, "Corpus building for data-driven TTS systems," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, 2002, pp. 199-202.