Disfluency Correction of Spontaneous Speech using Conditional Random Fields with Variable-Length Features

Jui-Feng Yeh[2], *Chung-Hsien Wu*[1] and *Wei-Yen Wu*[1]

[1] Department of Computer Science and Information Engineering, National Cheng Kung University
[2] Department of Computer Science and Information Engineering, National Chiayi University

jfyeh@csie.ncku.edu.tw, chwu@csie.ncku.edu.tw, wywu@csie.ncku.edu.tw

10.21437/Interspeech.2007-582

Abstract

This paper presents an approach to detecting and correcting edit disfluency based on conditional random fields with variable-length features. The variable-length features consist of word, chunk and sentence features. Conditional random fields (CRF) are adopted to model the properties of the edit disfluency, including repair, repetition and restart, for edit disfluency detection. For the evaluation of the proposed method, Mandarin conversational dialogue corpus (MCDC) is used. The detection error rate of edit word is 17.3%. Compared with DF-gram, Maximum Entropy and the approach combining language model and alignment model, the proposed approach achieves 11.7%, 8% and 3.9% improvements, respectively. The experimental results show that the proposed model outperforms other methods and efficiently detects and corrects edit disfluency in spontaneous speech.

Index Terms: Spontaneous speech processing, rich transcription, conditional random fields, disfluency correction

1. Introduction

The advanced information and computer science technologies have accomplished the commoditization that humans desire especially in abundance of persistent computation. However, a convenient user-oriented human machine interface is an essential issue for services provided by information technology. One of the most important human machine interfaces is speech-driven processing, especially in speech recognition. Recently, speech recognition technologies are close to maturity for read speech. However, the variances in speaking styles cause the drop of performance in spontaneous speech recognition. Since spontaneous speech is usually ill-formed, the conventional language model cannot achieve satisfactory improvement using the syntactic approaches such as parsing. Additionally, conventional n-gram model is unserviceable when the disfluencies, such as repair, repetition and restart, happen [1]. The edit difluencies in spontaneous speech should be treated as an important issue in recognition for practical applications such as conversational telephony system [2] .

According to the definition of disfluency, edit disfluencies can be categorized into tree types: repair, repetition and restart. The structure of an edit disfluency consists of deletable region, edit words, interruption point, editing term and correction part as shown in the Fig. 1.

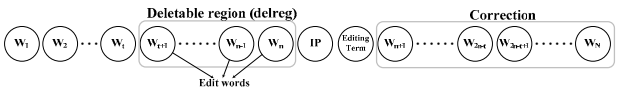

Figure 1: *The structure of an edit disfluency.*

Deletable region (delreg) means the speaker's initial attempt that exhibits some type of disfluency. The edit words means the words happen in the deletable region. The interruption point (IP) at which the speaker breaks off the deletable region with a repetition, repair or restart. Sometimes, editing term contains fillers or silence pause that occur within the context of an edit disfluency which is optional for all edit disfluencies. The correction part consists of the portion of the utterance that has been repaired by the speaker.

In the latest decade, a vast research effort had been invested in the edit disfluency detection and correction. Since 1998 ISCI and SRI International combined the language models and prosodic models to detect the disfluency [3]. Liu used the language models based on word and part-of-speech to address the problems resulted from repetition [4]. Taking edit disfluency as the hidden event in speech recognition process, the DF-gram models had been proposed [5]. Maximum entropy (ME), hidden Markov model (HMM) and conditional random fields were used to correct edit disfluency [6]. Bear et al. integrated multiple knowledge sources for detecting and correcting repairs in human computer dialog [7]. Stolcke et al. used the rule-based approach to providing an efficient repair procedure for quick transcriptions [8]. Honal and Schultz adopted the noisy channel model with different features to correct disfluency [9]. Charniak and Johnson built a part-of-speech based classifier to predict the deletable region [10]. Nakatani and Hirschberg proposed a decision tree solution integrating acoustic, prosodic and text-based cues to identify repairs in spontaneous speech [11]. Snover et al. and Kim et al. detected the disfluency by transformation-based learning [12-13]. Yeh and Wu integrated the alignment model and language model to build an edit disfluency cleanup model [14]. Tseng collected a Mandarin Conversational Dialogue Corpus (MCDC) for acoustic property analysis of disfluency [15]. Lin et al. adopted the maximum entropy model to detect interruption points (IP) of disfluent utterances.

Most of the previous research on edit disfluency either focused on a specific edit disfluency type or only considered single knowledge source. Lacking the integrated solution with needed resources, spontaneous speech recognition usually cannot be applied to conversational situation. Restated, the issue is how to detect and correct edit disfluency by combining the variable-length features from various resources. This paper proposes a novel disfluency cleanup model based on conditional random fields with variable-length features. Three variable-length units, word, chunk and sentence are employed as the states in the transition feature functions and

observation functions of the conditional random fields. Chunk is extracted by the Apriori algorithm [16] according to word co-occurrence. Sentence is identified according to the verb and its corresponding necessary arguments which appear in the neighbor in the sentence.

The rest of this paper is organized as follows. Section 2 describes the framework of the proposed cleanup model. Section 3 then illustrates the conditional random fields with variable-length features. Next, Section 4 presents the experiments to evaluate the proposed model. Finally, concluding remarks are made in Section 5.

2. The framework of proposed model

The framework of the proposed disfluency cleanup model is illustrated in Fig. 2.

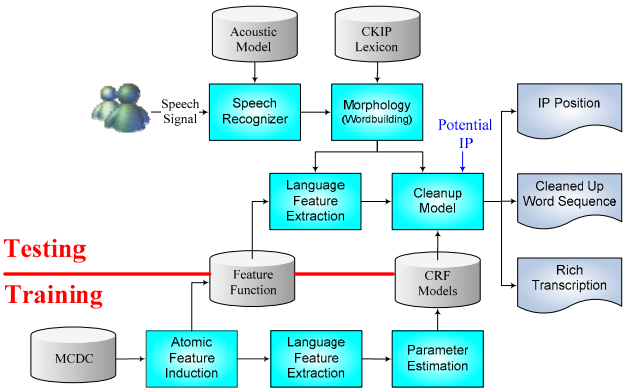

Figure 2: *Framework of the Disfluency Cleanup Model using Conditional Random Fields*

In this model, speech recognizer with 157 sub-syllable and 11 filler models, constructed by Hidden Markov Model Toolkit (HTK), is used to recognize the Mandarin speech input. Word-building procedure transforms the syllable lattice into the word lattice according to lexical information. The language-related features derived from Mandarin Conversational Dialogue Corpus (MCDC) by atomic feature induction module are fed to a feature function. Language feature extraction is further adopted to extract the language-related features from the word lattice. Based on the framework of conditional random fields (CRF), the improved iterative scaling (IIS) algorithm is adopted for estimating the weights to achieve the optimal performance. Finally, the cleanup model can detect and correct the disfluencies in the spontaneous speech input and provide the information about interruption point position, cleaned up word sequence and rich transcription.

3. Conditional random fields with variable length features

Conditional random fields (CRFs) are one of the state of the art approaches in information extraction taking advantage of sequence characteristics [17]. It has been used to solve the part of speech tagging problem and web searching problem and achieved commendably performance. In this paper, the cleanup of edit disfluency is regarded as a labeling problem. Cleanup is a process from the disfluent utterance to the corresponding fluent one. According to the structure of edit disfluency shown in Fig. 1, cleanup processing deletes the edit words that appear in the deletable region. Furthermore, we can label the words within the utterance as kept or deleted. For example, if the original utterance is "我(I) 明天(Tomorrow) * 明天(Tomorrow)) 要(will) 去(goto) 台北(Taipei)" and its corresponding fluent utterance is "我(I) 明天(Tomorrow)) 要(will) 去(goto) 台北(Taipei)." The star mark "*" means the interruption point and the first "明天(tomorrow)" in the original utterance is the edit word with repetition type and should be deleted. Since the goal is to label the words within the deletable region as "0", that is to say, they should be removed when the cleanup process is adopted. This example is illustrated in Fig. 3.

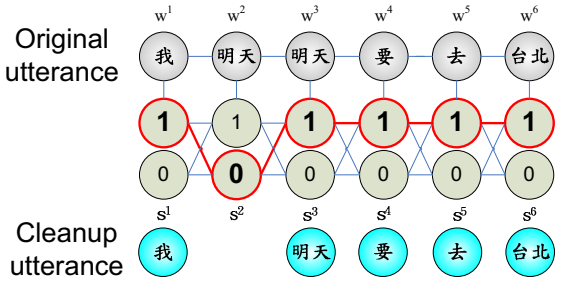

Figure 3: *One example presenting the process of word labeling in cleanup process.*

Since disfluency detection and correction can be simplified as labeling the edit word. The conditional random fields with the variable-length features are employed to detect the edit word. The labeling process can be formulated as:

$$\begin{aligned}\hat{\mathrm{S}} &= \arg\max_{\mathrm{S}} P(\mathrm{S}\,|\,\mathrm{X}) \\ &= \arg\max_{S}\left(\sum_{W} P(\mathrm{S}|\mathrm{W},\mathrm{X})\mathrm{P}(\mathrm{W}|\mathrm{X})\right) \\ &\cong \arg\max_{S}\left(\sum_{W} P(\mathrm{S}|\mathrm{W})\mathrm{P}(\mathrm{W}|\mathrm{X})\right)\end{aligned} \quad (1)$$

where X denotes the input speech signal; W means the word sequence from the speech recognition module, and S is the labeled state sequence.

There are two main probabilities computed in the right hand side of Eq.(1). Firstly, The term $\mathrm{P}(\mathrm{W}|\mathrm{X})$, which represents the posterior probability, is obtained from the speech recognizer and then $P(\mathrm{S}|\mathrm{W})$ is modeled by the conditional random fields as Eq.(2) shows.

$$P(S\,|\,W) = \frac{1}{Z}exp\left(\begin{array}{c}\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_q^{(t-1)}, s_p^{(t)}, \mathrm{W}\right) + \\ \sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)}, \mathrm{W}\right)\end{array}\right) \quad (2)$$

where $s_j^{(i)}$ denotes the *i-th* state and the *j-th* level feature. p and q represent the features used in the *t-th* and the (*t*-1)*-th* states. The features can be either word, chunk, or sentence. S is the state sequence that consists of the state $s_j^{(i)}$. $f_k(\cdot)$ and $g_k(\cdot)$ are the transition function and observation function, respectively. λ_k and μ_k are the weight parameters of the

observation function and transition function in the log linear combination model. Normalization factor Z is defined as the following:

$$Z = \sum_{W,s} \exp\left(\sum_{t} \sum_{k} \sum_{p,q} \lambda_k f_k\left(s_q^{(t-1)}, s_p^{(t)}, \mathrm{W}\right) + \sum_{t} \sum_{k} \sum_{p} \mu_k g_k\left(s_p^{(t)}, \mathrm{W}\right) \right) \quad (3)$$

Besides the word-level, chunk- and sentence-level features are included to form the variable-length feature set. Conditional random fields are used to model the edit disfluency in spontaneous speech as shown in Fig. 4.

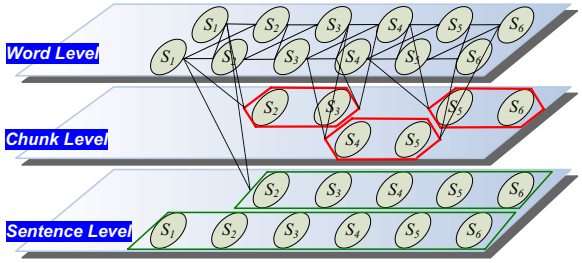

Figure 4: *Multiple level conditional random fields with word, chunk, and sentence level features.*

Compared to the conventional approaches, not only word level features but also the chunk and sentence level features are adopted as the input of the proposed model. In these features, word is a well-defined unit in speech and language processing. However, chunk and sentence are hard to be extracted from the speaker's utterance. Herein, the Apriori algorithm and verb-oriented sentence extraction algorithm are adopted to extract the chunk and sentence. In data mining, the Apriori algorithm is a classic algorithm for learning association rules according to co-occurrence. This study extracts the chunk using the Apriori algorithm as illustrated in Figure 5. The chunk patterns are formed from the item sets contained in the association rules by bottom-up subset exploration.

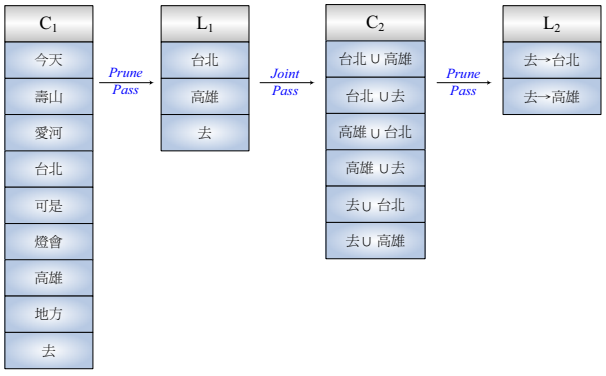

Figure 5: *The chunk pattern extraction using Apriori algorithm.*

The verb-oriented sentence extraction algorithm, an efficient and effective sentence extraction algorithm, is adopted to detect the sentence boundary. Academia Sinica has defined the verb's characteristic and their corresponding necessary arguments in [18]. The verb-oriented sentence extraction algorithm can be applied based on the verb's features. Firstly, the head word with the part-of-speech "verb" can be determined. Furthermore, the corresponding necessary arguments of the head word are detected. Finally, the sentence boundaries are decided by seeking the nearest necessary arguments from the head word. The example is presented in Figure 6.

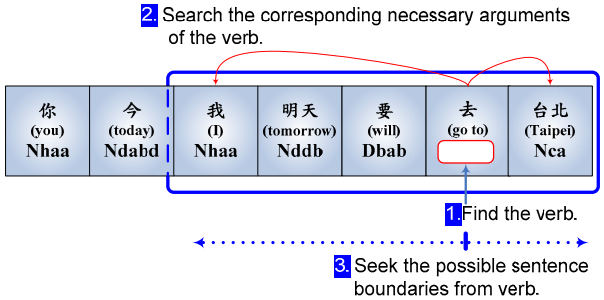

Figure 6: *The verb-oriented sentence extraction algorithm*

The improved iterative scaling algorithm (IIS) is employed to estimate the values of the parameters λ_k and μ_k. The weights are updated iteratively to obtain the optimal performance.

4. Experiments

The Mandarin conversational dialogue corpus (MCDC) [19] composed of 30 digital conversational dialogues with 27 hours was used to evaluate the proposed method in edit word detection of spontaneous speech. The edit word miss rate and false alarm rate were used to evaluate the performance of the proposed approach. Three edit disfluency types: repair, repetition and restart defined in SimpleMDE [20] (http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf) are considered as the observation events in this experiment.

The results of human generated transcription and speech recognition output are illustrated in Tables 1 and 2, respectively. The proposed method achieved significant improvement compared to DF-gram, maximum entropy (ME), hidden Markov model and the model combining tri-gram model and alignment model either in text data only or speech recognition result.

Table 1. *The performances of the methods using the human generated transcript.*

	Human generated transcription		
	Miss	False Alarm	Error
DF-gram	0.13	0.16	0.29
ME	0.05	0.20	0.25
HMM	0.12	0.14	0.26
3-gram+ alignment	0.09	0.12	0.21
Proposed	0.07	0.10	0.17

Table 2. *The performances of methods using the speech recognition output.*

	Speech recognition output		
	Miss	False Alarm	Error
DF-gram	0.37	0.346	0.71
ME	0.14	0.52	0.66
HMM	0.34	0.35	0.68
3-gram+ alignment	0.32	0.32	0.64
Proposed	0.25	0.35	0.60

Since interruption point is essential for edit word detection, Figure 7 shows the results of CRFs with variant feature set. The best result was obtained when the feature set contains word, chunk and sentence.

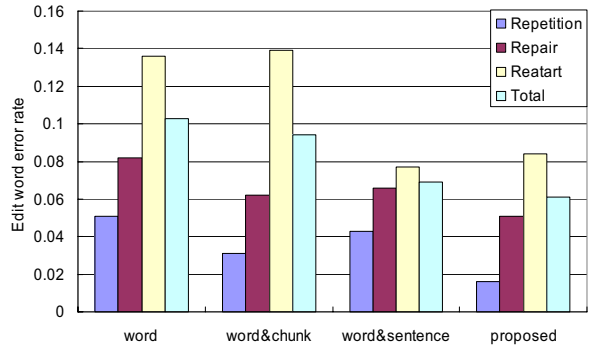

Figure 7: *The results of proposed model with variant feature sets given interruption point position.*

The performance of edit word detection gradually declines as the observation feature length increases. The reason is that the abandoned deteletable region usually contains few words and the correction part usually contains a whole sentence or sub-sentence. Another finding of this experiment is the unit using sentence can provide significant improvement on the resolution between "restart" and fluent sentence. Considering the characteristics of verbs and sentence structural information, we can achieve significant improvement on detecting the "restart" disfluency.

5. Conclusion

This paper has proposed a disfluency detection and correction model based on conditional random fields with variable length features. There are three features: word, chunk, and sentence. Based on the building block: word, this study extracted the chunk and sentence features using bottom-up subset exploration by the Apriori algorithm and verb-oriented sentence extraction algorithm, respectively.

According to the results of the experiments, the proposed method with the error rate for edit word detection is 17.3%. It achieved 11.7%, 8.7%, 8%, and 3% improvements compared to DF-Gram, hidden Markov model (HMM), maxmum entropy(ME) and the model combining the language model and alignment model. The above results show that the proposed method outperforms the conventional approaches in edit disfluency detection and correction.

6. References

[1] Kahn, J.-G., , M. Ostendorf and C. Chelba" Parsing Conversational Speech Using Enhanced Segmentation." Proc. HLT-NAACL, 2004. pp. 125-128.

[2] Soltau, H., , B. Kingsbury, , L. Mangu, , D. Povey, , G. Saon, , and D. Zweig, " The IBM 2004 Conversational Telephony System for Rich Transcription." In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05). (2005), 205-208.

[3] Stolcke, A., E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu." Automatic detection of sentence boundaries and disfluencies based on recognized words," In Proc. International Conference on Spoken Language Processing, pages 2247--2250, 1998.

[4] Liu, Y., E. Shriberg, and A. Stolcke. "Automatic disfluency identification in conversational speech using multiple knowledge sources," In Proc. Eurospeech, volume 1, pages 957—960, 2003.

[5] Stolcke, A. and E. Shriberg. "Statistical language modeling for speech disfluencies". In Proceedings of the International Conference of Acoustics, Speech, and Signal Processing, 1996.

[6] Liu, Y., E. Shriberg, A. Stolcke, M. Harper "Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection." Eurospeech 2005.

[7] Bear, J., J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detecting and correction of repairs in human computer dialog," in Proc. of ACL, 1992, pp. 56–63.

[8] Stolcke, A., W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in Proc. Intl. Conf. Spoken Language Processing, (Jeju, Korea), October 2004.

[9] Honal, M., and T. Schultz, "Correction of disfluencies in spontaneous speech using a noisy-channel approach," In EUROSPEECH-2003, 2781-2784.

[10] Charniak, E. and M. Johnson. "Edit detection and parsing for transcribed speech," In Proceedings of the North American Chapter of the Association for Computational Linguistics annual meeting, pages 118-126, 2001.

[11] Nakatani, C. and J. Hirschberg. "A corpus-based study of repair cues in spontaneous speech." Journal of the Acoustical Society of America, pages 1603-1616, 1994.

[12] Snover, M., B. Dorr, and R. Schwartz. "A lexically-driven algorithm for disfluency detection". In Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics annual meeting, 2004.

[13] Kim, J., S. E. Schwarm, and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning." Proceedings of HLT/NAACL 2004, pp. 137–144. 2004.

[14] Yeh, J.-F. and C.-H. Wu, "Edit Disfluency Detection and Correction Using a Cleanup Language Model and an Alignment Model," IEEE Trans. Audio, Speech, and Language Processing, 2006.

[15] Tseng, S.-C, "Repairs and Repetitions in Spontaneous Mandarin," In Proceedings of Workshop on Disfluency in Spontaneous Speech (DISS 03). Ed. Robert Eklund. Gothenburg Papers in Theoretical Linguistics 90. Pp. 71-74. University of Gothenburg.

[16] Chien, J.-T. " Association pattern Language Modeling," In IEEE Transaction on Audio, Speech, and Language Processing, Vol. 14, issue 5, 2006, pp.1719-1728

[17] Lafferty, J., McCallum, A., and Pereira, F. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." Proc. 18th International Conf. on Machine Learning., pp. 282-289. 2001.

[18] Academia Sinica, "Academia Sinica CKIP's Technical Report 93-05." 2004.

[19] Tseng, S.-C. and Liu, G.-L. "Annotation of Mandarin Conversational Dialogue Corpus." 2002.

[20] SimpleMDE http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf