Speech Reconstruction – Overview of State-of-the-art Systems

P. Češka

Charles University Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. This paper describes the speech reconstruction problem and lists the most frequent speaker errors. We show that deletion of these errors is not sufficient and more complex string operation should be done. We overview state-of-the-art methods which try to solve this problem and present new ideas which we would like implement in the near future.

Introduction

Spontaneous speech is the most natural form of language input. But majority of natural language processing (NLP) tools are designed for written language. They are trained, evaluated and applied only on the written text. The main reason behind this phenomenon is that spoken language lacks well-formedness and has unpredictable structure. In addition automatic speech recognition (ASR) transcripts include non-speech events like inhales, coughs and laughs. Task for speech reconstruction is to process ASR transcript to flawless, fluent and grammatically correct output which can be than processed by other NLP tools.

Rao et al. (2007) have shown that only simple disfluency removal can improve BLEU (standard evaluation metric for statistical machine translation) up to 8%. Cleaned transcripts also improve human readability as have been shown by Gibson et al. (2004).

Let us summarize the ongoing content: Right after the introduction follows a definition and analysis of speaker errors. In the next section we describe main linguistics resources for speech recognition (SR) and then we follow with description of state-of-the-art systems for speech recognition. Finally, we introduce our research ideas and the paper is summarized in the conclusion.

Defining speaker errors

Before we start describing speech reconstruction efforts we have to define what we exactly mean by speaker errors. Fitzgerald (2009) mentions that utterance is errorful if:

- It contains speaker self-repairs and disfluencies
- It is ungrammatical (disagreement in tense, number, or gender)
- It is incomplete, or
- It is inaccurately segmented into sentence-like unit

The most common disfluencies are classified as filler words (“*ah*”, “*um*”, “*eh*”), discourse markers (“*you know*”, “*I mean”)* and speaker edit regions. We recognize three different types of speaker edit regions:

- Repetition is part of the sentence which is repeated when the speaker stops for a while, thinking what to say next
- Revision occurs when speaker immediately corrects what he said before
- Restart fragment is the most complex type because reparandum is aborted by new thought

Figure 1 describes speaker edit region and its parts. For almost 80% of all speaker errors, deletion of reparandum and interregnum is satisfactory. But for some other problems more sophisticated transformations are needed. These problems are word reordering, rephrasing, non-grammatical sentences (no agreement between subject and predicate), sentence boundary errors, argument ellipsis and colloquial words (“*gonna,*” *“dobrej,” “štyři”*).

Linguistics resources for Speech Reconstruction

Before we start developing, training and testing speech reconstruction methods, we need corpus which will contain ASR transcriptions and reconstructed sentences created by annotators. There are four main sources of these data for English and Czech.

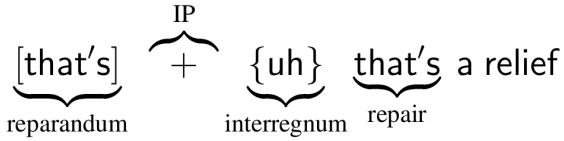

Figure 1. Typical edit region structure. This is example of repetition. The picture is taken from Fitzgerald (2009).

Switchboard (SWBD)

This corpus was prepared by Godfrey et al. (1992). It contains about 2500 conversations by 500 speakers from the United States. It's a well-known corpus for training and testing of a new speech algorithms. This corpus has been partially manually parsed and include a non-terminal node EDITED which indicates reparandum. New version of this corpus (released in PennTreebank 3) also has disfluency annotation.

Fisher Conversational Telephone Speech corpus (Fisher)

This corpus contains 16000 English conversations of duration more than 2000 hours. It's manually parsed like the Switchboard corpus. It has been created by Cieri et al. (2004).

Spontaneous Speech Reconstruction corpus (SSR)

This corpus has been build by Fitzgerald and Jelinek (2008). It's build atop Fisher corpus and is enriched by labeled word alignment between original and reconstructed utterances and semantic role labels for all verbs.

Prague Dependency Treebank of Spoken Language (PDTSL) – Hajič et al., 2008

This treebank contains data taken from Malach project and Companion project. They are mostly in Czech (more than 80 hours) but there are also some English conversations (more than 20 hours). All conversations are structured on 4 levels:

- audio, the lowest layer, contains only original audio recordings
- z-layer includes ASR output connected with word boundaries in audio layer
- w-layer is there for manual transcription of conversation and is connected with appropriate words of z-layer
- m-layer includes manually reconstructed transcript connected with m-layer

Possible techniques for Speech Reconstruction

This chapter will overview historical approaches on speech reconstruction and also discuss some other possible techniques borrowed from different NLP tasks. Almost all these techniques have been used only for English not Czech.

Early works narrow problem of speech reconstruction to detection and deletion of speaker edit regions. It's not surprise that first approaches relied on looking for identical sequences of words like Bear et al. (1992). Other early approaches relied also on acoustic and prosodic information (Nakatani and Hirschberg (1993)), statistical language model (Heeman and Allen (1999)) which includes identification of POS tags, discourse markers, speech repairs, and intonational phrases simultaneously. Also context-free grammar parser which includes non-terminal EDITED for identification of speaker edit regions has been implemented by Charniak (1999). Prediction of EDITED non-terminal as common non-terminals like NP or S doesn't work, probably due to inconsistency and infrequency of speaker edit regions.

We can also look at speech reconstruction from different angles. We can use paraphrasing and summarization techniques for deleting non-essential parts of utterances. If we imagine reconstruction as a translation between spoken and written language, then statistical machine translation techniques can be used. The most common approach for machine translation is a noisy channel model with this formula:

$$\hat{W} = \arg\max P(W \mid S) = \arg\max P(S \mid W)\, P(W)$$

where W stands for written language and S for spontaneous language. The probability P(W) is determined by language model. Standard approach for language model is to estimate probability on bases of a word sequence. A n-gram model approximate probability of string only by conditioning the previous n words. P(S | W) is translation model learnt from word alignment of paralel corpus. There are many translation toolkit which can be used for machine translation. The well-known basis is Moses toolkit created by Koehn et al. (2007).

State-of-the-art systems

We will introduce two state-of-the-art systems in this chapter. Firstly, we introduce model based on noisy channel paradigm, secondly we present conditional random field model which is the best performing system these days.

Noisy channel TAG model

Tree-adjoining grammar (TAG) is the channel model which recognizes very similar words in almost the same word order. It uses channel model paradigm:

$$\hat{W} = \arg\max P(W \mid S) = \arg\max P(S \mid W)\, P(W)$$

where W stands for written language and S for spontaneous language. A syntactic parser is used as the source model and a TAG based transducer is used as a channel model. TAG model is motivated by the intuition that the reparandum is a “rough copy” of the repair as describe Johnson and Charniak (2004). Most of probabilistic model assume that there are linear or tree-structured dependencies between reparandum and repair. Charniak and Johnson think that it seems to involve “crossed” dependencies and thus TAG model is appropriate as it's shown on Figure 2.

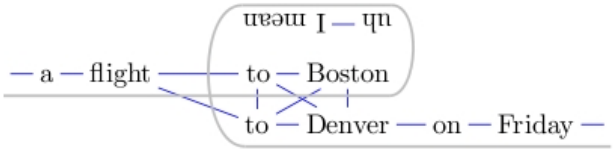

Figure 2. The “helical” dependency structure induced by the generative model of speech repairs. The picture is taken from Charniak and Johnson (2001).

We assume that W is a substring of S and can be obtained by deleting word from the spoken sentence. The TAG channel model represented way how to search the space of possible sentences W. Charniak used minimum edit-distance string alignment to learn TAG model probabilities (insertion and deletion is prefered before substitution).

Conditional random field and Maximum entropy prediction model

Conditional random fields (CRF) are undirected graphical models whose prediction of a sequence of hidden variables Y is globally conditioned on a given observation sequence X. Each observable state is composed of the corresponding word and set of additional features. More information can be found in Lafferty et al. (2001). This model has been successfully applied on many NLP tasks especially on tagging. The model has sequential context like Hidden Markov Model but it has less restricted feature set.

Fitzgerald (2009) choose these feature functions in her work:

- Lexical features – part-of-speech (POS) tag, token position within sentence (because many simple mistakes are at the beginning of utterances), etc.
- TAG based model – Johnson and Charniak model has been used as a feature
- Language model – token probability based on short history of previous tokens (for words and for POS)
- Non-terminal ancestors – the Charniak (1999) parser trained on SWBD corpus has been used. This feature is added because errors occur at certain point of phrases.

After evaluation of results with this model Fitzgerald pointed out that F-score (standard measure for tagging and also this NLP task) improves if the model is trained and tested only on errorful utterances. She divided the task to two steps – classify errorful utterances and than use CRF model to find errors.

Maximum entropy model is implemented for this purpose. It expanded previous identification using a maximum entropy model to combine previous feature model with these new features such as using deep linguistic parser to confirm well-formedness, unseen phrase rules expansion, utterance length, etc. This model is binary classifier for utterances and it significantly improve the results, especially the recall.

Research plans

We are now recording interviews between retired people and artificial being (Companions Project). These interviews will be added to PDTSL treebank to enrich interviews from Malach project. In the near future we will analyse speaker errors in Czech and compare them with errors in English (SSR corpus).

We will use the two step approach introduced by Fitzgerald – firstly find errorful utterances and secondly repair them. We will focus more on acoustic and prosodic features. We will implement pattern recognition model based on neural networks in cooperation with Marek Kukačka from Charles University in Prague, who already earlier used this approach for pattern recognition in pictures. We will use machine translation approach (more precisely Moses toolkit) for the second step of speech recognition problem.

Conclusion

In this paper we describe the task of SR, most common speaker error and we briefly introduce all methods which try to deal with this problem. Noisy channel TAG model and two step approach with maximum entropy model and conditional random field are described.

Our future work will be based on analysis of speaker errors in Czech language. We will develop prediction model of errorful utterances by pattern recognition on audio level and use statistical machine translation with Moses toolkit.

Acknowledgments. The work has been partially supported by the Czech Grant Agency under contract GA201/09/H057.

References

Bear, J., Dowding, J. And Shriberg, E. Integrating multiple knowledge sources for detection and crrection of repairs in human-computer dialog. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware, 1992.

Charniak, E. A maximum-entropy-inspired parser. In *Meeting of the North American Association for Computational Linguistics*, 1999.

Cieri, C., Strassel, S., Maamouri, M., Huang, S., Fiumara, J., Graff, D., Walker, K. and Liberman, M. Linguistic resource creation and distribution for EARS. In *Rich Transcription Fall Workshop*, 2004.

Companions Project, http://www.companions-project.org/.

Fitzgerald, E. and Jelinek F. Linguistic resources for reconstructing spontaneous speech text. In *Proceedings of the Language Resources and Evaluation Conference*, 2008.

Fitzgerald, E. Reconstructing spontaneous speech. Ph.D. thesis, The Johns Hopkins University, 2009.

Gibson, E., Wolf, F., Fedorenko, E., Jones, D., Chuang, C. and Patel, R. Two new experimental protocols for measuring speech transcript readability for timed question-answering task. In *Rich Transcription Fall Workshop*, 2004.

Godfrey, J. J., Holliman, E. C. and McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Internacional Conference on Acoustics, Speech and Signal Processing*, San Francisco, 1992.

Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptáček, J., Toman, J. and Urešová, Z. PDTSL: An anotated resource for speech reconstruction. In *Proceedings of the IEEE Workshop on Spoken Language Technology*, 2008.

Heeman, P. and Allen, J. Speech repairs, intonational phrases and discourse markers: Modeling speakers utterances in spoken dialogue. *Computational Linguistics*, 25(4), 1999.

Johnson, and Charniak, E. A TAG based noisy channel model of speech repairs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2004.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demonstration Session*, Prague, Czech Republic, 2007.

Lafferty, J., McCallum, A. and Pereira, F. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, San Francisco, USA, 2001.

MALACH Project, http://malach.umiacs.umd.edu/.

Nakatani, C. H. and Hirschberg, J. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America, 1993*.

Rao, S., Lane, I. and Schultz, T. Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts. In *Machine Translation Summit XI*, Copenhagen, Denmark, 2007.