Automatic disfluency detection in a dialogue system

Dana Dannélls
Department of Swedish Language
Göteborg University
SE-405 30 Gothenburg, Sweden
dana.dannells@svenska.gu.se

Abstract

Disfluency detection is the task of recognizing structural metadata in spoken utterances. It has been the topic of several studies in computational linguistics and psycholinguistics. This paper motivates the need for automatic disfluency detection in a dialogue system and delineates some of the features that characterize a disfluent utterance.

1 Introduction

Disfluency[1] detection is the task of recognizing structural metadata in spoken utterances such as fragmented words, self-corrections, stammering, repetitions, abandoned constituents, hesitations, and filled pauses. Disfluency detection is known to be a hard problem that is thought to require information on prosody, syntactic and semantic relations between constituents, discourse-level knowledge, and phonological well-formedness constraints. The awareness to this problem has started to gain momentum already during the early eighties (Hotopf, 1983; Fromkin, 1980).

Given how common disfluencies are in any type of discourse,[2] it seems natural that this phenomena is studied from multiple perspectives, for example it is studied in computational and psycholinguistics (for a detailed overview of these perspectives see Eklund, 2004). Here we study the phenomena from a computational linguistics perspective. While most of the work within the linguistic community have focused on detection and correction of disfluencies with emphasis on avoiding errors during the linguistic analysis process (Hirschberg et al., 2004), there are substantial work carried out from engineering perspective on constructing robust parsers capable of processing disfluent input. Considering the structure and goals of a dialogue system, it seems plausible to find solutions that enhance both the parser and the dialogue manager performance.

[1] The term *Disfluency* is defined in section 2.

[2] Disfluencies estimate to occur at the rate of about five to six per hundred words of spontaneous speech Bortfeld et al. (2001).

Disfluency detection in dialogue systems

Disfluencies signal misunderstanding, confusions, uncertainties and are an important source of information for dialogue systems. They have shown to be relevant at different levels of speech processing (Shriberg and Stolcke, 2004) as they: (1) cause recognition errors, and (2) indicate the user's intention in a dialogue. In this paper we focus on the later problem which is in particular relevant for a dialogue manager whose goal is to identify the intended speech act and decide on how to respond to this act and how to proceed with the dialogue.

Automatic disfluency detection for language understanding and speech recognition modules and the question of how to best structure interfaces that minimize disfluency in order to improve the robust performance of spoken dialogue systems are important issues that have drawn with it an intensive search for reliable methods to detect a disfluent input (Hindle, 1983; Nakatani and Hirschberg, 1994; O'shaughnessy, 1992; Shriberg, 1994). Disfluency detection could help spoken dialogue systems to facilitate user understanding by incrementally presenting the most relevant information and by adapting the interaction to address communication problems as they arise. A task which is feasible for dialogue systems (Jameson et al., 2001).

During an interaction with a dialogue system a communication problem may arise when a user is experiencing high cognitive load due to task complexity. In consequence, disfluencies such as unfilled pauses, self-corrections and repetitions increase (Müller et al., 2001). Within this research paradigm we investigate some of the features that characterize a disfluent utterance that could be used to detect and measure the individual level of cognitive load automatically.

The structure of the paper is as follows. In the next section we elaborate the notion of disfluency and motivate the need for automatic disfluency detection. In section 3 we present the techniques that can be utilized to accomplish this task. Section 4 delineates the acoustic and syntactic features that characterize disfluencies based on empirical findings. We end with conclusion and a few guidelines for future work.

2 Disfluency types and disfluency detection

The word "fluency" according to the *Oxford English dictionary* is:

> The quality or condition of being fluent, that is the ability to speak easily and accurately.

Disfluency is an antonym to fluency and especially with regards to speech can be defined as an inconsistent utterance in terms of grammatical structure and flow. Speech disfluency is an utterance where some of the words that the speaker utters needs to be removed in order to correctly understand the speaker's intention. An example of the structure of a speech disfluency is illustrated in Figure 1 (the figure is taken from Carletta et al., 1993).

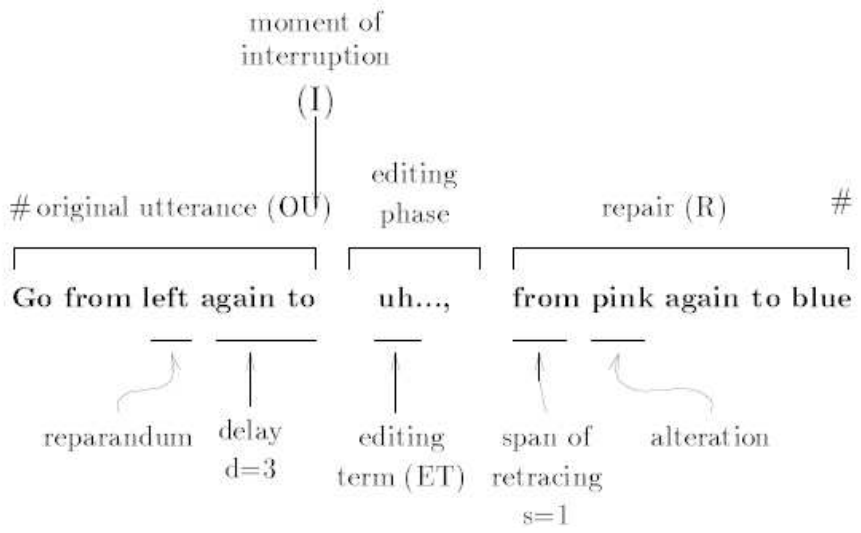

Figure 1: A structure of a speech disfluency

Speech disfluency can be divided into three types (Heeman and Allen, 1994):[3]

- fresh starts, when the speaker abandons what he/she says and starts again, e.g. *it was planed –*[4] *he started upp with an introduction*
- modifications, when the speaker modifies what he said before, e.g. *one way flight – one way fare*
- abridged, when the repair consists solely of a fragment and/or editing terms, e.g. *they want to – um get there sooner*

One motivation for automatic disfluency detection is the fact that disfluencies entail on-line planning, hesitation and self-correction, information which, among other things is crucial for a dialogue manager. Detection of the different disfluency types requires linguistic (but also non-linguistic) information which can be derived from corpus based studies. Corpus studies have been used to find the syntactic and prosodic regularities that characterize disfluencies and which can be utilized to model disfluency patterns. These studies provide insights into the nature of disfluencies in spoken dialogues and are used to examine speech variations and dialogue acts cross domains and under different circumstances (Nakatani and Hirschberg, 1994).

Another motivation for detecting disfluencies is the fact that a disfluent input can help in measuring the speaker's *cognitive load*. Cognitive load can be thought of as the mental

[3]Different authors use different divisions, and different appellation such as "restart","self-repair", etc. However they all refer to the same set of disfluency types.

[4]The symbol – denotes the right edge of the reparandum region and in this paper will be referred to as the interruption point (IP).

energy required to process a given amount of information. It can be increased by imposing a time pressure or increasing the difficulty of a task (Sweller, 1988). Today there are different techniques developed to measure a speaker cognitive load in which articulation rate measurements are observed.

There has been an increasing work in studying spontaneous speech, in both human-human and human-machine dialogs. In some of these studies disfluency is viewed as a signal that indicates the speaker is experiencing processing problems or experiencing high cognitive load (Clark and Wasow, 1998). Oviatt (1997) shows disfluencies are a sensitive predictor of planning demands and cognitive load during human-computer interaction.

3 Automatic disfluency detection

The task of automatic disfluency detection can be viewed as a binary statistical classification problem: Given a set of acoustic-prosodic and lexical-statistical features which are extracted from a recognized utterance, determine whether the input is disfluent. In combination with additional underlying factors, this information could help distinguishing whether the utterance signals cognitive load.

Machine learning and statistical modeling techniques have the advantage of being able to classify an utterance based on a number of potentially valuable features that describe that utterance and to determine which of the features are actually useful for obtaining a successful classification. Classifiers can be trained to determine when a user signals hesitation or distraction and are powerful approaches which have already been applied to a number of discourse problems with great success, including disfluency detection, dialogue act prediction, cue word usage and discourse segmentation.

Statistical and machine learning models are composed of a large number of instances, each of which contains an arbitrary number of features. The selected features have a large effect on the performance of the learning system and the choice of the features is therefore crucial for achieving good results.

4 Disfluency features

Features of speech characterize utterances and identify regularities in data, such as those manifested in disfluencies. These regularities include prosodic properties such as length, accent and stress, tone, intonation etc. Previous approaches to automatic disfluency detection have based their theories on task-oriented, domain-specific dialogues and proposed different features that characterize disfluencies for automatic disfluency detection. The proposed features are therefore based on combinations of lexical, semantic, syntactic and acoustic information which depend on the applied classification type and on the methods for measuring disfluency rates and the data sample. Below we present some of the features that characterize disfluency production according to previously tested theories and observations.

Prosodic features

Acoustic-prosodic features can be used to mark prosodic functions, i.e., prosodic boundary and phrase accent (Shriberg et al., 2000;Shriberg et al., 2003). Shriberg and Stolcke (2004) use prosodic features for automatic speech recognition and understanding. They describe how modeling of directly measurable prosodic features in combination with lexical-statistical language models can enhance accuracy on various tasks such as disfluency detection. They use phone-level alignment information which yields a rich inventory of features reflecting F0, pauses, segment duration and energy. Similarly, Sonmez et al. (1998) found disfluencies characterized by an increased F0 frequency and pause duration. They obtained F0 features by extracting pitch tracks from the speech signal which were post-processed to obtain stylized pitch contours. Pitch accents correspond roughly to pitch movements that lend emphasis to certain words in an utterance. Prosodic breaks are typically realized by a combination of a pause, a boundary-marking pitch movement, and lengthening of the phrase-final segments.

Heeman and Allen (1994), Nakatani and Hirschberg (1994) emphasis disfluency segment length increases in cases of high cognitive load. They point out duration values depend on the actual interruption point[5] of the word and phrase, and should therefore be taken into account on at least three levels, i.e., word duration, phrase duration and duration on phone-level. They further note phonetic alignments and phoneme segmentations have a large effect on the observed duration. Bell et al. (2003) have shown function words are less reduced and are longer in duration when they precede or follow disfluency.

According to Müller (2001) the vocal pitch seems to increase when an utterance is disfluent. Vowels are lengthened and consonants are shortened. These are therefore clues that might be helpful during the recognition process. This suggests that features such as duration of pauses, pitch of the last syllable in the utterance and information on whether it is stressed are important features to take into account. In addition to these features it is necessary to have information about the number of syllable per second that are produced by the speaker and the total duration value of an utterance.

Other prosodic clues that have been proven to be relevant for disfluency detection and in particular for the purpose of measuring high cognitive load are measurements of quantity. Additional prosodic features that have been proposed by previous authors and are relevant for automatic disfluency detection are: the disfluent word length (in case a word has been recognized); stressed syllables and the number of syllables in each word; duration of last syllable in the word; number of syllables and stressed syllables in the utterance; number of filled and unfilled pauses within the recognized utterance; total duration value; max word duration; mean phoneme duration, and amplitude mean value.

[5] An interruption point is marked with (I) in Figure 1.

Syntactic features

Liu et al. (2003) investigated the effect of word-language models in conjunction with acoustic-prosodic features and conclude that each knowledge source contributes differently to the combined performance. They found that taking repetition patterns into account in the language model leads to positive effect. This agrees with Oviatt (1995) who found longer utterances that consist of a large amount of words have higher disfluency rates than shorter utterances and that the quality of an utterance seem to decrease when it is disfluent.

Syntactic clues such as the grammatical structure of an utterance, its complexity and whether it is grammatically correct are indicators of disfluency. Syntax errors are often increased, especially under high cognitive load (Shriberg et al., 1992; Bear et al., 1992). As indicated by several authors (Bear et al., 1992; Heeman and Allen, 1994; O'shaughnessy, 1992; Nakatani and Hirschberg, 1994), part-of-speech cues of the target word and its surrounding context are strong indicators that can help predict whether an utterance signals disfluency. It has even been suggested (Oviatt, 1995; Hindle, 1983) that lexical information alone can be used to identify disfluencies. However, syntactic information depends heavily on the performance of the tagger, parser, chuncker etc., and part-of-speech tags may not always be reliable (Shriberg et al., 1997).

General features

Shriberg (2005) emphasizes the importance of modeling non-linguistic information to better understand the properties of natural speech. Lickley (1994) found that female speakers are less disfluent than male speakers. He also shows younger subjects are less disfluent and emphasis on the property of age as a significant feature. Another factor that plays an essential role in combination with prosodic features is whether the user is a native vs. non-native speaker. Additional non-linguistic features which seem to characterize disfluency are conventional role and turn-change (Bortfeld et al., 2001).

Other non-linguistic features that have been shown to help determine the speaker's cognitive load and may be relevant in combination with automatic disfluency detection include pulse, eye movement, pupil size etc. These features require techniques and knowledge which we will not discuss here.

5 Conclusion

Previous work on automatic disfluency detection has shown disfluencies follow certain patterns and regularities that could be incorporated into a comprehensive and predictive model for successful disfluency detection. As noted, the task involves access to various features as pitch accents, prosodic boundaries at different locations, etc. Guided on previous findings we delineated a number of features which are relevant for detecting a disfluent utterance and which are suitable for machine learning techniques. Based on empirical evidence and theoretical findings we believe that computing linguistic and non linguistic features simul-

taneously is a powerful approach that incorporation with other attributes could help to determine whether an utterance is disfluent and whether it signals cognitive load. This will in turn provide a dialog manager with a vital piece information that can help in deciding on the dialog act. More knowledge about the nature of disfluency is further required in order to increase confidence in detecting it.

It will be interesting to test whether empirical evidences from domain specific dialogue corpus studies in combination with powerful tools such as machine learning algorithms and high-level statistical modeling actually improve the dialogue management. In the aspect of that the user himself will have a more pleasant interaction with the dialogue system as it will resemble a human-human dialogue. Some interesting questions, ought to be tested, are how well disfluency recognition improves a dialogue system with respect to the dialogue system architecture and most importantly, how can features such as presented in this paper be measured automatically.

References

A. Batliner, A. Buckow, H. Niemann, E. Nóth, and V. Warnke. The prosody module. In *Verbmobil: Foundations of Speech-to-Speech Translations*, pages 106–121, 2000. Springer, Berlin.

A. Batliner, A. Buckow, H. Niemann, E.Nóth, and V. Warnke. How to find trouble in communication. In *Speech Communication*, volume 40, pages 117–143, 2003.

J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics, 1992.

A. Bell, D.Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *Journal of the Acoustical Society of America (JASA)*, 113(2):1001–1024, 2003.

H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44:123–147, 2001.

J. Carletta, R. Caley, and S. Isard. A collection of self-repairs from the map task corpus, 1993.

H. H. Clark and T. Wasow. Repeating words in spontaneous speech. In *Cognitive Psychology*, volume 37, pages 204–242, 1998.

R. Eklund. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. PhD thesis, Department of Computer and Information Science Linkóping Studies in Science and Technology, 2004.

V. A. Fromkin. *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand.* Academic Press: New York, 1980.

P. Heeman and J. Allen. Detecting and correcting speech repairs. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 295–302, Morristown, NJ, USA, 1994. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/981732.981773.

D. Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, 1983.

J. Hirschberg, D. Litman, and M. Swerts. Prosodic and other cue to speech recognition failure. In *Speech Communication*, volume 4, pages 155–175, 2004.

W. H. N. Hotopf. Lexical slips of the pen and tongue: What they tell us about language production. In *Development, Writing and Other Language Processes*, volume 2, 1983.

A. Jameson, B. Großmann-Hutter, L. March, R. Rummer, T. Bohnenberger, and F. Wittig. When actions have consequences: Empirically based decision making for intelligent user interfaces. *Knowledge-Based Systems*, 14:75–92, 2001.

R. J. Lickley. *Detecting disfluency in spontaneous speech.* PhD thesis, University of Edinburg, 1994.

Y. Liu, E. Shriberg, and A. Stolcke. Automatic disfluency identification in conversational speech using multiple knowledge sources, 2003.

C. Müller, B. Grossmann-Hutter A. Jameson, R. Rummer, and F. Wittig. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In *User Modeling: Proceedings of the Eighth International Conference.* Springer, Berlin, 2001.

C. Nakatani and J. Hirschberg. A corpus based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 3:1603–1616, 1994.

D. O'shaughnessy. Analysis and automatic recognition of false starts in spontaneous speech. In *International Conference on Spoken Language Processing*, pages 931–934, 1992. Banff, Alberta.

S. L. Oviatt. Predicting spoken disfluencies during human-computer interaction. In *Computer Speech and Language*, pages 9–15, 1995.

S. L. Oviatt. Multimodal interactive maps: Designing for human performance. human computer interaction. In *Human Computer Interaction*, pages 93–129, 1997.

E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies.* PhD thesis, University of California at Berkeley, 1994.

E. Shriberg. Spontaneous speech: How people really talk and why engineers should care. In *Eurospeech*, 2005.

E. Shriberg and A. Stolcke. Prosody modeling for automatic speech recognition and understanding. In *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 139–146, 2004. Red Bank, New Jersey, USA.

E. Shriberg, J. Bear, and J. Dowding. Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 23–26, 1992. Harriman.

E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for. disfluency detection. In *Eurospeech*, 1997.

K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proceedings of the ICSLP*, pages 3189–3192, 1998.

J. Sweller. Cognitive load during problem-solving:effects on learning. In *Cognitive Science*, volume 12, 1988.