Disfluency Detection with a Semi-Markov Model and Prosodic Features

James Ferguson, Greg Durrett and **Dan Klein**
Computer Science Division
University of California, Berkeley
{jferguson,gdurrett,klein}@berkeley.edu

Abstract

We present a discriminative model for detecting disfluencies in spoken language transcripts. Structurally, our model is a semi-Markov conditional random field with features targeting characteristics unique to speech repairs. This gives a significant performance improvement over standard chain-structured CRFs that have been employed in past work. We then incorporate prosodic features over silences and relative word duration into our semi-CRF model, resulting in further performance gains; moreover, these features are not easily replaced by discrete prosodic indicators such as ToBI breaks. Our final system, the semi-CRF with prosodic information, achieves an F-score of 85.4, which is 1.3 F_1 better than the best prior reported F-score on this dataset.

1 Introduction

Spoken language is fundamentally different from written language in that it contains frequent disfluencies, or parts of an utterance that are corrected by the speaker. Removing these disfluencies is desirable in order to clean the input for use in downstream NLP tasks. However, automatically identifying disfluencies is challenging for a number of reasons. First, disfluencies are a syntactic phenomenon, but defy standard context-free parsing models due to their parallel substructures (Johnson and Charniak, 2004), causing researchers to employ other approaches such as pipelines of sequence models (Qian and Liu, 2013) or incremental syntactic systems (Honnibal and Johnson, 2014). Second, human processing of spoken language is complex and mixes acoustic and syntactic indicators (Cutler et al., 1997), so an automatic system must employ features targeting all levels of the perceptual stack to achieve high performance. In spite of this, the primary thread of work in the NLP community has focused on identifying disfluencies based only on lexicosyntactic cues (Heeman and Allen, 1994; Charniak and Johnson, 2001; Snover et al., 2004; Rasooli and Tetreault, 2013). A separate line of work has therefore attempted to build systems that leverage prosody as well as lexical information (Shriberg et al., 1997; Liu et al., 2003; Kim et al., 2004; Liu et al., 2006), though often with mixed success.

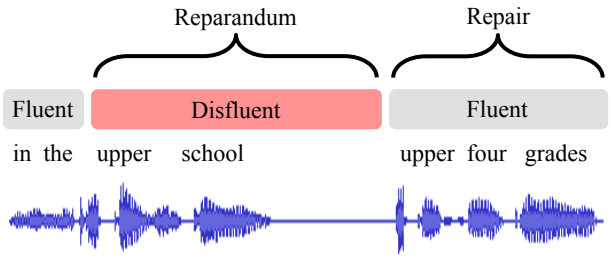

Figure 1: Example of a disfluency where the speaker corrected *upper school*. Our model considers both transcribed text and the acoustic signal and predicts disfluencies as complete chunks using a semi-Markov conditional random field.

In this work, we present a model for disfluency detection that improves upon model structures used in past work and leverages additional prosodic information. Our model is a semi-Markov conditional random field that distinguishes disfluent chunks (to be deleted) from fluent chunks (everything else), as shown in Figure 1. By making chunk-level predictions, we can incorporate not only standard token-level features but also features that can consider the entire reparandum and the start of the repair, enabling our model to easily capture parallelism between these two parts of the utterance.[1] This frame-

[1]The reparandum and repair are important concepts that we will refer to in this paper, but the model does not distinguish the repair from other fluent text which follows.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 257–262,
Denver, Colorado, May 31 – June 5, 2015.

work also enables novel prosodic features that compute pauses and word duration based on alignments to the speech signal itself, allowing the model to capture acoustic cues like pauses and hesitations that have proven useful for disfluency detection in earlier work (Shriberg et al., 1997). Such information has been exploited by NLP systems in the past via ToBI break indices (Silverman et al., 1992), a mid-level prosodic abstraction that might be indicative of disfluencies. These have been incorporated into syntactic parsers with some success (Kahn et al., 2005; Dreyer and Shafran, 2007; Huang and Harper, 2010), but we find that using features on predicted breaks is ineffective compared to directly using acoustic indicators.

Our implementation of a baseline CRF model already achieves results comparable to those of a high-performance system based on pipelined inference (Qian and Liu, 2013). Our semi-CRF with span features improves on this, and adding prosodic indicators gives additional gains. Our final system gets an F-score of 85.4, which is 1.3 F_1 better than the best prior reported F-score on this dataset (Honnibal and Johnson, 2014).

2 Experimental Setup

Throughout this work, we make use of the Switchboard corpus using the train/test splits specified by Johnson and Charniak (2004) and used in other work. We use the provided transcripts and gold alignments between the text and the speech signal. We follow the same preprocessing regimen as past work: we remove partial words, punctuation, and capitalization to make the input more realistic.[2] Finally, we use predicted POS tags from the Berkeley parser (Petrov et al., 2006) trained on Switchboard.

3 Model

Past work on disfluency detection has employed CRFs to predict disfluencies using a IOBES tag set (Qian and Liu, 2013). An example of this is shown in Figure 2. One major shortcoming of this model is that beginning and ending of a disfluency are not decided jointly: because features in the CRF are local to emissions and transitions, features in this model cannot recognize that a proposed disfluency begins with *upper* and ends before another occurrence of *upper* (see Figure 1). Identifying instances of this parallelism is key to accurately predicting disfluencies. Past work has captured information about repeats using token-level features (Qian and Liu, 2013), but these still apply to either the beginning or ending of a disfluency in isolation. Such features are naturally less effective on longer disfluencies as well, and roughly 15% of tokens occurring in disfluencies are in disfluencies of length 5 or greater. The presence of these longer disfluencies suggests using a more powerful semi-CRF model as we describe in the next section.

3.1 Semi-CRF Model

The model that we propose in this work is a semi-Markov conditional random field (Sarawagi and Cohen, 2004). Given a sentence $x = (x_1, \ldots, x_n)$ the model considers sequences of labeled spans $\bar{s} = ((\ell_1, b_1, e_1), (\ell_2, b_2, e_2), \ldots, (\ell_k, b_k, e_k))$, where $\ell_i \in \{\text{Fluent}, \text{Disfluent}\}$ is a label for each span and $b_i, e_i \in \{0, 1 \ldots n\}$ are fenceposts for each span such that $b_i < e_i$ and $e_i = b_{i+1}$. The model places distributions over these sequences given the sentence as follows:

$$p_\theta(\bar{s}|x) \propto \exp\left(\theta^\top \sum_{i=1}^{k} f(x, (\ell_i, b_i, e_i))\right) \quad (1)$$

where f is a feature function that computes features for a span given the input sentence. In our model we constrain the transitions so that fluent spans can only be followed by disfluent spans. For this task, the spans we are predicting correspond directly to the reparanda of disfluencies, since these are the parts of the input sentences that should be removed. Note that our feature function can jointly inspect both the beginning and ending of the disfluency; we will describe the features of this form more specifically in Section 3.2.2.

To train our model, we maximize conditional log likelihood of the training data augmented with a loss function via softmax-margin (Gimpel and Smith, 2010). Specifically, during training, we maximize $\mathcal{L}(\theta) = \sum_{i=1}^{d} \log p'_\theta(\bar{s}|x)$, where $p'_\theta(\bar{s}|x) = p_\theta(\bar{s}|x) \exp(\ell(\bar{s}, \bar{s}^*))$. We take the loss function

[2] As described in Honnibal and Johnson (2014), we computed features over sentences with filler words (*um* and *uh*) and the phrases *I mean* and *you know* removed.

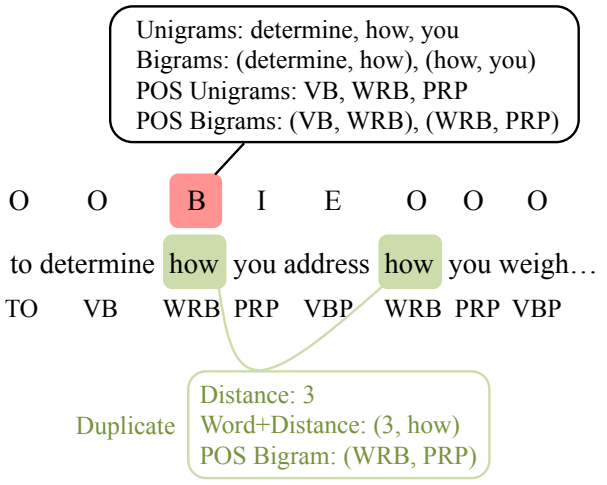

Figure 2: Token features for CRF and semi-CRF models.

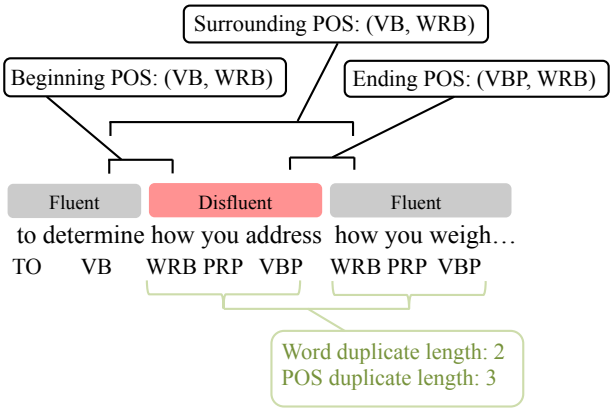

Figure 3: Span features for semi-CRF model.

ℓ to be token-level asymmetric Hamming distance (where the output is viewed as binary edited/non-edited). We optimize with the AdaGrad algorithm of Duchi et al. (2011) with L_2 regularization.

3.2 Features

Features in our semi-CRF factor over spans, which cover the reparandum of a proposed disfluency, and thus generally end at the beginning of the repair. This means that they can look at information throughout the reparandum as well as the repair by looking at content following the span. Many of our features are inspired by those in Qian and Liu (2013) and Honnibal and Johnson (2014). We use a combination of features that are fired for each token within a span, and features that consider properties of the span as a whole.

3.2.1 Token Features

Figure 2 depicts the token-level word features we employ in both our basic CRF and our semi-CRF models. Similar to standard sequence modeling tasks, we fire word and predicted part-of-speech unigrams and bigrams in a window around the current token. In addition, we fire features on repeated words and part-of-speech tags in order to capture the fact that the repair is typically a partial copy of the reparandum, with possibly a word or two switched out. Specifically, we fire features on the distance to any duplicate words or parts-of-speech in a window around the current token, conjoined with the word identity itself or its POS tag (see the *Duplicate* box in Figure 2). We also fire similar features for POS tags since substituted words in the repair frequently have the same tag (compare *address* and *weigh*). Finally, we include a duplicate bigram feature that fires if the bigram formed from the current and next words is repeated later on. When this happens, we fire an indicator for the POS tags of the bigram. In Figure 2, this feature is fired for the word *how* because *how you* is repeated later on, and contains the POS tag bigram (*WRB, PRP*).

Table 1 shows the results for using these features in a CRF model run on the development set.[3]

3.2.2 Span Features

In addition to features that fire for each individual token, the semi-CRF model allows for the inclusion of features that look at characteristics of the proposed span as a whole, allowing us to consider the repair directly by firing features targeting the words following the span. These are shown in Figure 3. Critically, repeated sequences of words and parts-of-speech are now featurized in a coordinated way, making it less likely that spurious repeated content will cause the model to falsely posit a disfluency.

We first fire an indicator of whether or not the entire proposed span is later repeated, conjoined with the length of the span. Because many disfluencies

[3]We created our development set by randomly sampling documents from the training set. Compared to the development set of Johnson and Charniak (2004), this more closely matches the disfluency distribution of the corpus: their development set has 0.53 disfluent tokens per sentence, while our set has 0.38 per sentence, and the training set has 0.37 per sentence.

	Prec.	Rec.	F_1
CRF	84.0	82.1	83.0
Semi-CRF	88.6	81.7	85.0
Semi-CRF + Prosody	89.5	82.7	86.0

Table 1: Disfluency results on the development set. Adding span features on top of a CRF baseline improves performance, and including raw acoustic information gives further performance gains.

are just repeated phrases, and longer phrases are generally not repeated verbatim in fluent language, this feature is a strong indicator of disfluencies when it fires on longer spans. For similar reasons, we fire features for the length of the longest repeated sequences of words and POS tags (the bottom box in Figure 3). In addition to general repeated words, we fire a separate feature for the number of uncommon words (appearing less than 50 times in the training data) contained in the span that are repeated later in the sentence; consider *upper* from Figure 1, which would be unlikely to be repeated on its own as compared to stopwords. Lastly, we include features on the POS tag bigrams surrounding each span boundary (top of Figure 3), as well as the bigram formed from the POS tags immediately before and after the span. These features aim to capture the idea that a disfluency is a mistake with a disjuncture before the repair, so the ending bigram will generally not be a commonly seen fluent pair, and the POS tags surrounding the reparandum should be fluent if the reparandum were removed.

Table 1 shows that the additional features enabled by the CRF significantly improve performance on top of the basic CRF model.

4 Exploiting Acoustic Information

Section 3 discussed a primarily structural improvement to disfluency detection. Henceforth, we will use the semi-CRF model exclusively and discuss two methods of incorporating acoustic duration information that might be predictive of disfluencies. Our results will show that features targeting raw acoustic properties of the signal (Section 4.1) are quite effective, while using ToBI breaks as a discrete indicator to import the same information does not give benefits (Section 4.2)

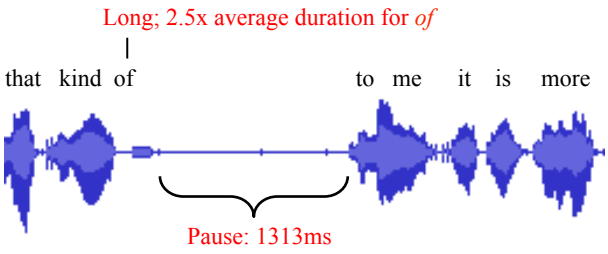

Figure 4: Raw acoustic features. The combination of a long pause and considerably longer than average duration for *of* is a strong indicator of a disfluency.

4.1 Raw Acoustic Features

The first way we implemented this information was in the form of raw prosodic features related to pauses between words and word duration. To compute these features, we make use of the alignment between the speech signal and the raw text. Pauses are then simply identified by looking for pairs of words whose alignments are not flush. The specific features used are indicators of the existence of a pause immediately before or after a span, and the total number of pauses contained within a span. Word duration is computed based on the deviation of a word's length from its average length averaged over all occurrences in the corpus.[4] We fire duration features similar to the pause features, namely indicators of whether the duration of the first and last words in a span deviate beyond some threshold from the average, and the total number of such deviations within a span. As displayed in Table 1, adding these raw features results in improved performance on top of the gains from the semi-CRF model.

4.2 ToBI Features

In addition to the raw acoustic features, we also tried utilizing discrete indicators of acoustic information, specifically ToBI break indices (Silverman et al., 1992). Previous work has shown performance improvements resulting from the use of such discrete information in other tasks, such as parsing (Kahn et al., 2005; Dreyer and Shafran, 2007; Huang and Harper, 2010). We chose to focus specifically on ToBI breaks rather than on ToBI tones because tonal information has appeared relatively less

[4]Note that this averages over multiple speakers as well.

	Disfluency		
	Prec.	Rec.	F_1
Baseline	88.61	81.69	85.01
AuToBI *3*, *4*	88.46	81.92	85.06
CRF ToBI	88.42	81.96	85.07
Raw acoustic	89.53	82.74	**86.00**

Table 2: Disfluency results with predicted ToBI features on the development set. We compare our baseline semi-CRF system (Baseline) with systems that incorporate prosody via predictions from the AuToBI system of Rosenberg (2010) and from our CRF ToBI predictor, as well as the full system using raw acoustic features.

	Prec.	Rec.	F_1
Johnson and Charniak (2004)	–	–	79.7
Qian and Liu (2013)	–	–	83.7
Honnibal and Johnson (2014)	–	–	84.1
CRF	88.7	78.8	83.4
Semi-CRF	90.1	80.0	84.8
Semi-CRF + Prosody	90.0	81.2	85.4

Table 3: Disfluency prediction results on the test set; our base system outperforms that of Honnibal and Johnson (2014), a state-of-the-art system on this dataset, and incorporating prosody further improves performance.

useful for this task (Shriberg et al., 1997). Moreover, the ToBI break specification (Hirschberg and Beckman, 1992) stipulates a category for strong disjuncture with a pause (*2*) as well as a pause marker (*p*), both of which correlate well with disfluencies on gold-annotated ToBI data.

To investigate whether this correlation translates into a performance improvement for a disfluency detection system like ours, we add features targeting ToBI annotations as follows: for each word in a proposed disfluent span, we fire a feature indicating the break index on the fencepost following that word, conjoined with where that word is in the span (beginning, middle, or end).

We try two different ways of generating the break indices used by these features. The first is using the AuToBI system of Rosenberg (2010), a state-of-the-art automatic ToBI prediction systems based on acoustic information which focuses particularly on detecting occurrences of *3* and *4*. Second, we use the subset of Switchboard labeled with ToBI breaks (Taylor et al., 2003) to train a CRF-based ToBI predictor. This model employs both acoustic and lexical features, which are both useful for ToBI prediction despite breaks being a seemingly more acoustic phenomenon (Rosenberg, 2010). The acoustic indicators that we use are similar to the ones described in Section 4 and our lexical features consist of a set of standard surface features similar to those used in Section 3.2.1.

In Table 2 we see that neither source of predicted ToBI breaks does much to improve performance. In particular, the gains from using raw acoustic features are substantially greater despite the fact that the predictions were made in part using similar raw acoustic features. This is somewhat surprising, since intuitively, ToBI should be capturing information very similar to what pauses and word durations capture, particularly when it is predicted based partially on these phenomena. However, our learned ToBI predictor only gets roughly 50 F_1 on break prediction, so ToBI prediction is clearly a hard task even with sophisticated features. The fact that ToBI cannot be derived from acoustic features also indicates that it may draw on information posterior to signal processing, such as syntactic and semantic cues. Finally, pauses are also simply more prevalent in the data than ToBI markers of interest: there are roughly 40,000 pauses on the ToBI-annotated subset of the dataset, yet there are fewer than 10,000 *2* or *p* break indices. The ToBI predictor is therefore trained to ignore information that may be relevant for disfluency detection.

5 Results and Conclusion

Table 3 shows results on the Switchboard test set. Our final system substantially outperforms the results of prior work, and we see that this is a result of both incorporating span features via a semi-CRF as well as incorporating prosodic indicators.

Acknowledgments

This work was partially supported by BBN under DARPA contract HR0011-12-C-0014 and by a Facebook Fellowship for the second author. Thanks to the anonymous reviewers for their helpful comments.

References

Eugene Charniak and Mark Johnson. 2001. Edit Detection and Parsing for Transcribed Speech. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Anne Cutler, Delphine Dahan, and Wilma van Donselaar. 1997. Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, 40(2):141–201.

Markus Dreyer and Izhak Shafran. 2007. Exploiting Prosody for PCFGs with Latent Annotations. In *Proceedings of Interspeech*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, July.

Kevin Gimpel and Noah A. Smith. 2010. Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions. In *Proceedings of the North American Chapter for the Association for Computational Linguistics*.

Peter Heeman and James Allen. 1994. Detecting and Correcting Speech Repairs. In *Proceedings of the Association for Computational Linguistics*.

Julia Hirschberg and Mary E. Beckman. 1992. The tobi annotation conventions. Online at http://www.cs.columbia.edu/ julia/files/conv.pdf.

Matthew Honnibal and Mark Johnson. 2014. Joint Incremental Disfluency Detection and Dependency Parsing. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pages 131–142.

Zhongqiang Huang and Mary Harper. 2010. Appropriately Handled Prosodic Breaks Help PCFG Parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Mark Johnson and Eugene Charniak. 2004. A TAG-based Noisy-channel Model of Speech Repairs. In *Proceedings of the Association for Computational Linguistics*.

Jeremy G. Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective Use of Prosody in Parsing Conversational Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Joungbum Kim, Sarah E Schwarm, and Mari Ostendorf. 2004. Detecting Structural Metadata with Decision Trees and Transformation-Based Learning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2003. Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources. In *Proceedings of Eurospeech*.

Yang Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. 2006. Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. *Transactions of Audio, Speech and Language Processing*, 14(5):1526–1540, September.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the Conference on Computational Linguistics and the Association for Computational Linguistics*.

Xian Qian and Yang Liu. 2013. Disfluency Detection Using Multi-step Stacked Learning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Mohammad Sadegh Rasooli and Joel Tetreault. 2013. Joint Parsing and Disfluency Detection in Linear Time. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Andrew Rosenberg. 2010. AuToBI - A Tool for Automatic ToBI Annotation. In *Proceedings of Interspeech*.

Sunita Sarawagi and William W. Cohen. 2004. Semi-Markov Conditional Random Fields for Information Extraction. In *Proceedings of Advances in Neural Information Processing Systems*.

Elizabeth Shriberg, Rebecca Bates, and Andreas Stolcke. 1997. A Prosody-only Decision-tree Model for Disfluency Detection. In *Proceedings of Eurospeech*.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A Standard for Labeling English Prosody. In *Proceedings of the International Conference on Spoken Language Processing*.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2004. A Lexically-Driven Algorithm for Disfluency Detection. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: An overview.