Unediting: Detecting Disfluencies Without Careful Transcripts

Victoria Zayats, Mari Ostendorf and Hannaneh Hajishirzi
Electrical Engineering Department
University of Washington
Seattle, WA, USA
[vzayats, ostendor, hannaneh]@u.washington.edu

Abstract

Speech transcripts often only capture semantic content, omitting disfluencies that can be useful for analyzing social dynamics of a discussion. This work describes steps in building a model that can recover a large fraction of locations where disfluencies were present, by transforming carefully annotated text to match the standard transcription style, introducing a two-stage model for handling different types of disfluencies, and applying semi-supervised learning. Experiments show improvement in disfluency detection on Supreme Court oral arguments, nearly 23% improvement in F1.

1 Introduction

Many hearings, lectures, news broadcasts and other spoken proceedings are hand-transcribed and made available online for easier searching and increased accessability. For speed and cost reasons, standard transcription services aim at representing semantic content only; thus, filled pauses (uh, um) and many disfluencies (repetitions and self corrections) are omitted, though not all. Careful transcripts represent all the words (and word fragments spoken), as shown below with disfluent regions underlined.

Careful: *It is it is a we submit*
Where there used to be um um um uh the decision
Standard: *It is, it is, we submit*
Where there used to be the decision

These phenomena are quite common in spontaneous speech, even in formal settings such as Supreme Court oral arguments and congressional hearings (Zayats et al., 2014).

While disfluencies may not be important for analyzing the topic of a discussion, the rate and type of disfluencies provide an indication of other factors of interest in spoken language analysis, including cognitive load, emotion, and social cues (Shriberg, 2001). Further, predicting locations of disfluencies in standard transcripts would help to improve time alignments of transcripts to the audio signal, and to provide more useful text data for training language models for speech recognition. Since careful annotation of transcripts with this information is costly, this paper tackles the problem of recovering the disfluencies from clues in the standard orthographic transcripts, or "unediting" the transcripts.[1]

Here, unediting is treated as detection of the reparandum of the disfluencies. Following the structural representation of (Shriberg, 1994), as in:

[we would + which we would]
[would + [who + who] wouldn't]

the task is to detect the words in the brackets preceding the '+' which marks the self-interruption point. Of course, here, some of the words in those regions may not be in the transcript, so location is more important than extent. In addition, some cues used (i.e. filled pauses and word fragments) are not available in standard transcripts.

Three developments are combined to address the problem of unediting with the constraint of limited hand-annotated training data in the target domain: oral arguments from the Supreme Court of the United States (SCOTUS) available from the Oyez Project archive (oyez.org). First, we identify mechanisms for transforming the careful transcripts of the Switchboard corpus (Godfrey et al., 1992) to be

[1] Thanks to Mark Liberman for the term "unediting."

more similar to the Oyez transcripts. Second, we introduce a multi-stage model that accounts for differences in the rates of repetitions and self-corrections in standard vs. careful transcripts. Lastly, we apply semi-supervised learning to take advantage of the large amount of original Oyez transcripts. The system combining all these techniques, referred to here as UNEDITOR, leads to an improvement in F1 of nearly 23% compared to a baseline of training from the original disfluency-annotated Switchboard corpus.

2 Related work

This paper builds on prior work using conditional random field (CRF) models (Liu et al., 2006; Georgila, 2009; Ostendorf and Hahn, 2013; Zayats et al., 2014). More recent work has shown a benefit from Markov networks (Qian and Liu, 2013; Wang et al., 2014). Since our work is on the transcription style mismatch, this work adopts the simpler CRF approach, but can be easily extended to other classification techniques.

In this work, we use only text features. While prosodic features have been shown to be useful (Shriberg, 1999; Kahn et al., 2005; Liu et al., 2006; Wang et al., 2014), the fact that the Oyez transcripts do not capture all the words means that forced time alignments are unreliable and the associated prosodic features are too noisy to be useful. Other studies integrate disfluency detection with parsing, e.g. (Charniak and Johnson, 2001; Johnson and Charniak, 2004; Lease et al., 2006; Hale et al., 2006; Miller, 2009; Miller et al., 2009; Zwarts et al., 2010; Rasooli and Tetreault, 2013; Honnibal and Johnson, 2014), but parsers trained on standard treebank data sets are not effective on the very long and complex sentences in SCOTUS; parser adaptation is left for future work.

There are a few studies that have investigated disfluency detection using cross-domain training data (Georgila et al., 2010; Ostendorf and Hahn, 2013; Zayats et al., 2014), and many more that have used multi-domain data for other language processing tasks. What is different about the task addressed here is that both the domain (topic and speaking style) and the transcription protocol differ between the target and source domain. There have been some attempts to transform written text to a more conversational style for training language models, e.g. Bulyko et al. (2007) inserted pause fillers and word repetitions, which led to reductions in perplexity though not word error rate. The work here differs in that the transformation is in the reverse direction (removing fillers from conversational text) and punctuation cues are emphasized.

3 Transforming training data

Here we describe methods for generating training data for use with standard transcripts: i) transferring labels from a small amount of carefully annotated data to corresponding standard transcripts, and ii) transforming the existing Switchboard training set to make it more similar to the target domain.

3.1 SCOTUS corpora

The Oyez Project at Chicago-Kent is a multimedia archive containing audio and transcripts of the Supreme Court hearings since 1955. While OYEZ transcripts are consistent with the audio in general, they are not accurate when it comes to disfluencies. We notice that most simple disfluencies such as repetitions have been omitted by OYEZ annotators, while more complex ones are often present and annotators have used the '...' symbol at locations of filled pauses or repetitions. Having those explicit cues indicating interruption points in disfluencies makes it possible to consider recovering the untranscribed disfluencies.

For CAREFUL SCOTUS annotation, we use the data provided by (Zayats et al., 2014), which includes seven cases with carefully transcribed audio and hand-annotated disfluencies, with separately marked repetitions. We develop ANNOTATED OYEZ transcripts, by transferring disfluency labels for those seven cases from CAREFUL SCOTUS to the corresponding files in OYEZ and dropping the deletion markers. As a result, those transcripts are identical to the original OYEZ transcripts, but in addition contain disfluency annotation derived from CAREFUL SCOTUS.

In order to align the CAREFUL SCOTUS and ORIGINAL OYEZ transcripts, we use a dynamic programming algorithm for sequence alignment with matching scores as given in Table 1 and a deletion

CAREFUL SCOTUS	OYEZ	Score
exact match	exact match	4
'+'	'...'	3
punctuation	punctuation	2
end of sentence	'...'	2
word/punct	'...'	1
word	other word	-1
word	punct	-1

Table 1: Matching scores used in dynamic programming transcript alignment.

cost of 1. Some examples of CAREFUL SCOTUS, OYEZ, ALIGNED OYEZ (with deletions marked) and ANNOTATED OYEZ transcripts are shown below. The full corpus is available at `https://ssli.ee.washington.edu/tial/data/oyez`.

CAREFUL SCOTUS: *[[S It is + it is] a +] we submit*
Where there used to be um um um uh the decision
OYEZ: *It is, it is, we submit*
Where there used to be the decision
ALIGNED OYEZ: *[[S It is, + it is], _ +] we submit*
Where there used to be _ _ _ _ the decision
ANNOTATED OYEZ: *[[S It is, + it is], +] we submit*
Where there used to be the decision

3.2 Switchboard transformation

The ANNOTATED OYEZ training set is a very small dataset, and other work has shown that Switchboard (SWBD) is useful for cross-domain training for SCOTUS (Zayats et al., 2014). However, prior work has been with careful transcripts. SWBD transcripts do not include '...' symbols, and SWBD has many more commas and other punctuation symbols. In order to make best use of the SWBD data, we transform it to be more similar to the OYEZ transcripts in two steps. First, we add '...' after interruption points in SWBD. Second, we remove all punctuations except '...' in the middle of the sentence in both of the corpora.

4 Detecting disfluencies

In this section we describe the UNEDITOR system, which is a two-stage CRF model trained on transformed training data and takes advantage of a large pool of unlabeled data with a self-training technique.

Baseline: CRF We use a conditional random field (CRF) model that labels each word in a sentence, following a tagging approach with separate repetition and non-repetition reparandum states, as in (Ostendorf and Hahn, 2013). The feature set includes identity and pattern match features widely used in disfluency detection tasks, as well as distance-based and disfluency language model features from (Zayats et al., 2014).

4.1 Two-stage model

Using the same features as in the baseline, we introduce a two-stage CRF model motivated by our observation that many repetitions are omitted from the standard transcriptions. Thus, while 62% of disfluencies in CAREFUL SCOTUS are repetitions, only 22% of all disfluencies in ANNOTATED OYEZ are repetitions. We find that training at two separate stages helps to overcome the difference in distributions of two disfluency types between source and target domains, and hence results in a better model for adaptation. In the first stage, we train a model to detect repetitions by only considering repetition states in the training data. In the second stage, we train a model to detect non-repetitions by removing all repetitions from the training data. Similarly at test time, we use the first-stage model to detect repetitions, then remove all the detected repetitions, and apply the second-stage model to detect non-repetitions. In evaluation, we report the disfluencies detected in both stages.

4.2 Self-training

A benefit of OYEZ transcripts is that there is a huge amount of unlabeled data available, which makes it natural to use semi-supervised learning. In this work, we use a simple self-training approach. First we apply a CRF model trained on the labeled data to the unlabeled data. Then we augment the training data with automatically labeled sentences that have been detected to contain a disfluency with a confidence score greater than 0.5, and retrain the model with the new augmented training set.

5 Experiments and discussion

We evaluate the different sources/transformations of training data, self-training and the two-stage detection model on ANNOTATED OYEZ transcripts from three cases ($\sim$30k words).

Training set	Prec	Rec	F1
CAREFUL SCOTUS	66.1	16.7	26.7
ANNOT OYEZ	**86.7**	20.4	33.0
ORIG SWBD	62.2	29.1	39.7
CAREFUL SCOTUS + ORIG SWBD	63.7	27.8	38.7
ANNOT OYEZ + TRANSF SWBD	70.9	**49.0**	**57.9**

Table 2: Disfluency detection of ANNOT OYEZ with different training sets.

Training set: ANNOT OYEZ+	Prec	Rec	F1
ORIGSWBD	67.8	29.3	40.9
SWBD WITH ...	63.1	46.8	53.7
TRANSF SWBD	**70.9**	**49.0**	**57.9**

Table 3: The combination of ANNOT OYEZ and SWBD with different SWBD transformation steps.

Training set	Prec	Rec	F1
CAREFUL SCOTUS	57.8	21.2	31.0
ANNOT OYEZ	**81.7**	27.3	41.0
ORIG SWBD	59.0	31.7	41.2
CAREFUL SCOTUS + ORIG SWBD	64.6	33.7	44.3
ANNOT OYEZ + TRANSF SWBD	71.7	**52.8**	**60.8**

Table 4: Self-training performance using different initial models.

5.1 Transforming training data

First, we assess the utility of different training sources and training data transformation using the baseline model. Note that the two SCOTUS sets are quite small (four cases, ∼64k words) compared to Switchboard (1.3M words). Because of the difference in punctuation style between the original Oyez transcripts and the careful transcripts of both corpora, all sentence-internal punctuation is removed in the CAREFUL SCOTUS and ORIG SWBD data.

Table 2 reports results on training the CRF model with the different sources and their combinations. As expected, detection with in-domain training data and transformed SWBD (ANNOT OYEZ+TRANSF SWBD) outperforms training on all other dataset combinations. Training on ANNOT OYEZ alone significantly outperforms detection (especially precision) when only trained on the carefully annotated data because of the matching transcription style. Training with ORIG SWBD outperforms training with ANNOT OYEZ alone mainly due to the availability of more training data in the SWBD dataset, consistent with results in (Ostendorf and Hahn, 2013). Surprisingly, the CAREFUL SCOTUS data did not provide any benefit when added to the ORIG SWBD.

Next, we study the impact of adding '...' symbols and removing punctuation for transforming the SWBD data. Table 3 reports results for training the CRF model with the combination of ANNOT OYEZ and SWBD with different transformation steps. We observe that roughly 30% of the interruption points in CAREFUL SCOTUS are associated with the '...' symbol in the OYEZ transcripts; therefore, we add '...' symbols after $1/3$ of the interruption points in the SWBD. As expected, disfluency detection is improved by transforming SWBD with adding '...'. The largest gain is obtained when we also remove punctuation (the row TRANSF SWBD). All further experiments use this setting for training the models.

5.2 Self-training

Here we study the contribution of semi-supervised learning when applied on the baseline model (Table 5). For self-training, we use 1,765 OYEZ transcripts dated 1990 - 2011 as our unlabeled data (∼17.5M words), with a confidence threshold of 0.5 for augmenting the training data, as described previously. We use each one of the baseline models in Table 2 as an initial model for the self-training for comparison to the results in Table 4. While adding a lot of in-domain data definitely helps, the quality of the initial model plays a major role in the overall performance.

5.3 Two-stage model

Finally, we assess the impact of the two-stage model with and without self-training (Table 5). For the two-stage semi-supervised model, self-training was only used for the second stage (non-repetition detection). As expected, both two-stage and self-training models improve the baseline CRF model, and the combination performs the best. The two-stage model helps to adapt the differences in distribution of repetitions and non-repetitions between the two domains by factoring the different problems to improve the match of the more difficult non-repetition cases. Overall, we obtain nearly 23% im-

Model	Prec	Rec	F1
1-stage	70.9	49.0	57.9
1-stage semi	71.7	**52.8**	60.8
2-stage	**83.3**	47.6	60.6
UNEDITOR: 2-stage semi	76.8	52.2	**62.2**

Table 5: Baseline, two-stages and self-training methods, comparison: baseline self-training method is trained on ...., all the rest methods are trained on ANNOT OYEZ and TRANSF SWBD. Our method, UNEDITOR combines self-training and two-stage models.

provement using the full UNEDITOR system comparing to the model trained on the ORIG SWBD dataset.

6 Conclusion

In this paper we present a framework for disfluency detection in non-careful transcripts. Experiments are based on the OYEZ archive of transcriptions of Supreme Court oral arguments. To address the problem of lack of annotated data, we first transfer disfluency annotations from careful transcripts of a few cases to the less precise OYEZ transcripts. Next, we transform Switchboard transcripts to make them more similar to the target domain. In addition, we introduce a two-stage model and self-training to further improve performance.

Experiments show improvement in disfluency detection on Supreme Court oral arguments. Starting from baselines of training from carefully annotated in-domain data (F1=26.1) or Switchboard data (F1=39.7), we achieve a substantial improvement to (F1=62.2) with our best case system UNEDITOR, which corresponds to an improvement of nearly 23% over the stronger baseline.

Possible extensions of this work include exploring graph-based semi-supervised approaches (e.g., (Subramanya et al., 2010)) and combining the text-based approach with flexible ASR forced alignment allowing optional insertion of filled pauses and words that are common as repetitions. In addition, the availability of the automatically annotated disfluencies makes it possible to study the variation in rates for different cases and speakers over an extended time period.

Acknowledgments

This work was supported in part by DARPA grant FA8750-12-2-0347. The authors thank the anonymous reviewers for their valuable feedback to improve the clarity of paper. The authors also thank Sangyun Hahn for his contribution in the two-stage model. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

[Bulyko et al.2007] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and O. Cetin. 2007. Web resources for language modeling in conversational speech recognition. *IEEE-TSLP*, 5(1).

[Charniak and Johnson2001] E. Charniak and M. Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proc. NAACL*, pages 118–126.

[Georgila et al.2010] K. Georgila, N. Wang, and J. Gratch. 2010. Cross-domain speech disfluency detection. In *Proc. Annual SIGdial Meeting on Discourse and Dialogue*.

[Georgila2009] K. Georgila. 2009. Using integer linear programming for detecting speech disfluencies. In *Proc. NAACL-HLT*.

[Godfrey et al.1992] J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proc. ACL*, volume I, pages 517–520.

[Hale et al.2006] John Hale, Izhak Shafran, Lisa Yung, Bonnie Dorr, Mary Harper, Anna Krasnyanskaya, Matthew Lease, Yang Liu, Brian Roark, Matthew Snover, and Robin Stewart. 2006. PCFGs with syntactic and prosodic indicators of speech repairs. In *Proc. COLING-ACL*.

[Honnibal and Johnson2014] Matthew Honnibal and Mark Johnson. 2014. Joint incremental disfluency detection and dependency parsing. *TACL*, 2(1):131–142.

[Johnson and Charniak2004] M. Johnson and E. Charniak. 2004. A tag-based noisy channel model of speech repairs. In *Proc. ACL*.

[Kahn et al.2005] Jeremy G Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective use of prosody in parsing conversational speech. In *Proc. EMNLP-HLT*, pages 233–240.

[Lease et al.2006] Matthew Lease, Mark Johnson, and Eugene Charniak. 2006. Recognizing disfluencies in conversational speech. *IEEE-TASLP*, 14(5):169–177.

[Liu et al.2006] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE-TASLP*, 14:1526–1540.

[Miller et al.2009] Tim Miller, Luan Nguyen, and William Schuler. 2009. Parsing speech repair without specialized grammar symbols. In *Proc. ACL-IJCNLP*, pages 277–280.

[Miller2009] Tim Miller. 2009. Improved syntactic models for parsing speech with repairs. In *Proc. NAACL-HLT*.

[Ostendorf and Hahn2013] Mari Ostendorf and Sangyun Hahn. 2013. A sequential repetition model for improved disfluency detection. In *Proc. Interspeech*.

[Qian and Liu2013] Xian Qian and Yang Liu. 2013. Disuency detection using multi-step stacked learning. In *Proc. NAACL-HLT*.

[Rasooli and Tetreault2013] Mohammad Sadegh Rasooli and Joel R Tetreault. 2013. Joint parsing and disfluency detection in linear time. In *Proc. EMNLP*, pages 124–129.

[Shriberg1994] E. Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Department of Psychology, University of California, Berkeley, CA.

[Shriberg1999] E. Shriberg. 1999. Phonetic consequences of speech disfluency. In *Proc. ICPhS*, pages 619–622.

[Shriberg2001] Elizabeth Shriberg. 2001. To errrris human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(01):153–169.

[Subramanya et al.2010] Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proc. EMNLP*, pages 167–176. ACL.

[Wang et al.2014] Xuancong Wang, Hwee Tou Ng, and Khe Chai Sim. 2014. A beam-search decoder for disfluency detection. In *Proc. COLING*.

[Zayats et al.2014] Victoria Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2014. Multidomain disfluency and repair detection. In *Proc. Interspeech*.

[Zwarts et al.2010] Simon Zwarts, Mark Johnson, and Robert Dale. 2010. Detecting speech repairs incrementally using a noisy channel approach. In *Proc. COLING*, pages 1371–1378.