Linguistic Resources for Reconstructing Spontaneous Speech Text

Erin Fitzgerald, Frederick Jelinek

Center for Language & Speech Processing, Johns Hopkins University
3400 N. Charles St., Baltimore, Maryland 21218
{erinf, jelinek}@jhu.edu

Abstract

The output of a speech recognition system is not always ideal for subsequent downstream processing, in part because speakers themselves often make mistakes. A system would accomplish speech reconstruction of its spontaneous speech input if its output were to represent, in flawless, fluent, and content-preserving English, the message that the speaker intended to convey. These cleaner speech transcripts would allow for more accurate language processing as needed for NLP tasks such as machine translation and conversation summarization, which often rely on grammatical input. Recognizing that supervised statistical methods to identify and transform ill-formed areas of the transcript will require richly labeled resources, we have built the Spontaneous Speech Reconstruction corpus. This small corpus of reconstructed and aligned conversational telephone speech transcriptions for the Fisher conversational telephone speech corpus (Strassel and Walker, 2004) was annotated on several levels including string transformations and predicate-argument structure, and will be shared with the linguistic research community.

1. Introduction

The output of a speech recognition system is often not what is required for subsequent processing, in part because speakers themselves often make mistakes (e.g. stuttering, self-correcting, or using filler words). A cleaner speech transcript would allow for more accurate language processing as needed for natural language processing tasks such as machine translation and conversation summarization which often assume a grammatical sentence as input. A system would accomplish speech reconstruction of its spontaneous speech input if its output were to represent, in flawless, fluent, and content-preserving English, the message that the speaker intended to convey.

Transforming errorful text using supervised statistical methods requires a gold-standard corpus of manually reconstructed sentences, which prior to this effort has never been produced. Anticipating the training and evaluation needs ahead as research in this area progresses, we produced a small corpus of reconstructed and aligned telephone speech text annotated on several levels including string transformations and predicate-argument structure, referred to as the Spontaneous Speech Reconstruction (SSR) corpus. Additional instances of text enrichment, such as adding capitalization and punctuation as appropriate, was considered to be outside the scope of this work.

2. Resource Motivation

While some annotated corpora have previously been produced for related problems, we believe that a need exists for expanded linguistic resources before automatic cleaning and transforming speech transcripts without altering the original content can accurately be done. The SSR corpus aims to fill this role.

2.1. Reconstructing Spontaneous Speech

The most similar existing language resource was produced by the Linguistic Data Consortium (LDC) in preparation for the 2004 NIST Rich Transcription Metadata Extraction (MDE) task on the Fisher conversational telephone speech (CTS) corpus (Cieri et al., 2004; Strassel and Walker, 2004). The goals of this task included accurate sentence segmentation and identification of simple disfluencies like **filler words** (i.e. "um", "ah", discourse markers ("you know"), and **edit regions** consisting of a *reparandum*, an *interruption point (IP)*, an optional *interregnum* (like "I mean"), and a *repair* region (Shriberg, 1994), as seen in Figure 1.

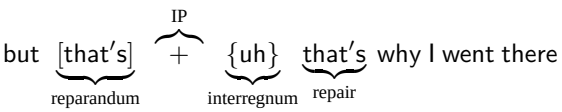


Figure 1: Typical edit region structure.

While the MDE-labeled Fisher corpus is a useful starting point for speaker error identification, the errors labeled are limited to a few types, and no recommendations are made as to how to fix the errors; simply deleting the identified reparandum regions is not always optimal. The SSR corpus annotated in this work builds on the LDC effort, using the same speech utterances (the Fisher CTS corpus) and giving the annotators access to LDC disfluency labels, but going a step further such that corrections (including potential insertions, substitutions, and constituent moves) with labeled alignments are recommended, and deeper predicate-argument analysis of the resultant reconstructions is also provided, as described in Section 4.

Examples of speaker errors which go beyond the simple edit region formalism include

1. still wants to party
   becomes
   [ARG] still wants to party

2. [how can you get that without] + it's a catch-22
   becomes
   how can you get that without [ARG] || it's a catch-22

3. i actually working in new jersey
   becomes
   i am actually working in new jersey

4. I haven't saw the old one but I saw new one
   becomes
   I haven't seen the old one but I saw the new one

5. they like video games and stuff some kids do
   becomes
   some kids do like video games and stuff
   or
   some kids like video games

In the above reconstructions, examples 1, 2, 3, and 4 involve word insertions, examples 2 and 5 preserve parts of the traditional reparandum region, example 2 divides a single sentence into two, example 4 involves a substitution, and example 5 requires coreference identification and phrase movement. Example 5 also demonstrates that there can be ambiguity in determining the best possible reconstruction of a given sentence.

2.2. Predicate-Argument Labeling for Reconstructed Speech Text

Every English verb is associated with a set of mandatory and optional argument roles, sometimes called a **roleset**. For example, the verb "say" must have a *sayer* and an *utterance which is said*, along with an optionally defined *hearer* and any number of locative, temporal, manner, etc. adjunctival arguments.

The task of predicate-argument labeling (sometimes called semantic role labeling or SRL) assigns this simple *who* did *what* to *whom when*, *where*, *why*, *how*, etc. structure to sentences, often for downstream processes such as information extraction and question answering. Reliably identifying and assigning these roles to grammatical text like news text is an active area of research (Gildea and Jurafsky, 2002; Pradhan et al., 2004), using training resources like the Linguistic Data Consortium's Proposition Bank (Palmer et al., 2005), a 300k-word corpus with semantic role relations labeled for verbs in the Wall Street Journal section of the Penn Treebank.

With an appropriately annotated conversational text training corpus we believe that these methods can be adapted for transcriptions of spontaneous speech as well in future research, and have incorporated these annotations into the SSR corpus. Rather than attempting to label incomplete utterances or errorful phrases, our annotators labeled sentences which were well-formed post-reconstruction. We believe the transitive bridge between the original and reconstructed sentences and reconstructions with their predicate-argument labels may yield insight into the structure of speech errors and how to extract these verb-argument relationships in verbatim speech text.

Furthermore, given a set of semantic role labels on an ungrammatical string, and armed with the knowledge of a set of core semantosyntactic principles which constrain the set of possible grammatical sentences, we hope to discover and take advantage of new cues for construction errors in the field of speech reconstruction.

3. Extracting a densely errorful corpus

Before investing time and resources repairing speech segments, it is to our advantage to first attempt to identify which utterances are poorly constructed (defined as being ungrammatical, incomplete, or missing necessary sentence boundaries prior to reconstruction). Extracting these sentences allows us to produce a densely errorful data set for effective training and efficient annotation.

We implemented several approaches to automatically identify these sentences. To evaluate the methods, we randomly sampled 500 sentences from our dataset and annotated each sentence s in the sample S as "good" or "bad", forming the set of poorly constructed sentences $P \subset S$. We then considered several approaches for utterance-level identification of the poor constructions P. Of these approaches, we found that the union of the Johnson and Charniak (2004) simple disfluency detection system – where all sentences with identified edits were labeled as poor – and a deep linguistic (Head-driven Phrase Structure Grammar (Pollard and Sag, 1994; Callmeier, 2001)) parser output – where all sentences not parsed were considered ill-formed – yielded the best overall sentence-level identification results, with 80.6% precision and 87.9% recall, as seen in Table 1. Accordingly, it was this combination which was used to extract data to be considered for annotation from the Fisher development and evaluation subcorpora, pruning the 21,456 sentences of the subcorpora into 6,384 utterances likely to contain errors to be manually reconstructed by trained annotators.

4. Building an Annotated Reconstruction Corpus

In a four-month effort, we trained annotators to reconstruct approximately 6,400 sentences (prefiltered as described in Section 3.) from the Fisher conversational telephone speech corpus for use as training and evaluation sets for future reconstruction endeavors.

SSR annotations were recorded via a custom-built tool shown in Figure 2 and described in Section 4.1., which was capable of storing and labeling the many types of changes we anticipated during the course of sentence-level reconstruction.

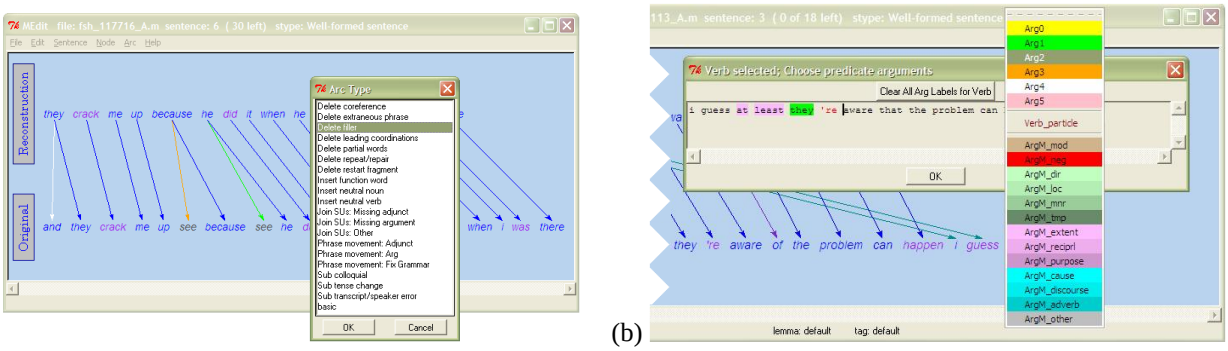
Because any given ill-formed sentence may have several valid reconstructions (as demonstrated in example 5 of Section 2.1.), each sentence was reconstructed independently by two or three annotators. This yielded two major benefits: we were able to better evaluate and ensure annotation quality, and the released data could contain multiple reconstructions to allow for more flexible task evaluation during the course of research. In all cases, annotators were encouraged to make the simplest changes necessary to make the sentence clean and grammatical with minimal change to its meaning.

Each sentence was annotated on three levels:

- **Word and alignment level.** The words in each utterance were deleted, inserted, substituted, or moved as required to make the sentence as grammatical as possible without altering the original meaning and without the benefit of context information. Sentences could

Approach	"Poor" Sentence	P	R	F
Edit detection	$\lvert\{\text{edits in } s\}\rvert \geq 1$	96.0	73.5	83.3
HPSG Parsable	s is parsable	78.7	67.4	72.6
Edits + HPSG	Approach 2 $\cup$ Approach 5	80.6	87.9	84.1

Table 1: Comparision of poor construction identification approaches on 500-sentence sample S.



(a) (b)

Figure 2: (a) Reconstruction example and (b) predicate-argument labeling example as viewed by the annotation tool.

also be split into two as required. Alignments between the original and reconstructed word sequences were defined, and for each alteration a corresponding label was chosen to explain the change made.

- **Utterance level.** Once reconstruction was complete, the final state of each reconstruction was manually labeled with one of six levels of grammaticality:
  - Well-formed and grammatical sentence
  - Well-formed fragment with content (ex. "Last June" or "Why not?")
  - Fragment without content (ex. "and it uh")
  - Backchannel Acknowledgement (ex. "Uh-huh" or "Sure")
  - Cannot repair utterance
- **Predicate-Argument Structure labeling.** For every well-formed and grammatical sentence (and only those utterances), all non-auxiliary verbs were identified and the corresponding predicate-argument structure was labeled according to the role-sets defined in the LDC Proposition Bank annotation effort (Palmer et al., 2005).

4.1. Developing the Annotation Tool

To accomplish efficient and consistent annotation, we aimed to build an annotation tool that was task-specific, simple to use (even for annotators with little linguistic training), and capable of storing and labeling the many types of changes we anticipated during the course of a given sentence-level reconstruction. We adapted a prototype annotation tool, designed at our partner institution Charles University in Prague for labeling simple Czech disfluencies of the type shown in Figure 1. Our revised tool had an expanded set of allowable word change types and included capabilities for semantic role argument labeling of verbs as reconstruction features (see Section 2.2.), and for semantic role labeling work.

The tool features separate modes for sentence reconstruction and predicate-argument labeling of reconstruction output, as well as a summary screen for reviewing annotation work accomplished at a glance. Our tool initially displays each original sentence linked word-by-word to a duplicate of the same sentence to be reconstructed. Annotators had access to the original audio files to help reduce interpretational ambiguity, and were able to correct many types of errors through the following set of reconstruction actions.

- **Delete** words: fillers, repetitions/revisions, restarts, coreference, leading conjugation
- **Insert** neutral elements: (ex. the, is, or an undefined noun phrase placeholder
- **Substitute** words: change tense or number, transcriber errors, colloquial phrases
- **Move** words within sentence boundaries: adjuncts, arguments, other grammar-necessary reorderings
- **Add sentence boundaries** to split sentences
- **Remove sentence boundaries** to adjoin consecutive sentences
- **Align** all original words with their "source" word(s) in the reconstruction (i.e. in the noisy channel paradigm)
- **Label** all changes and their rationale to track the problems identified and for training to reproduce these types of transformations
- **Label** the state of the final reconstruction: fragment, clean sentence, unable to repair, etc.
- **Identify** all active verbs in well-formed sentences, and label all primary and adjunct arguments

4.2. Annotation Characteristics and Statistics

Examining the reconstruction annotations produced, and moreover various agreement statistics between annotators reconstructing the same sentence, it becomes obvious how much variance exists in the set of valid reconstructions for a given sentence. A set of agreement statistics can be reviewed in Table 2.

Statistic	% pairwise matching
Exact string match	57%
Sentence-type match	86%
Word match	94%
Word count match	63%
Average Rec-Rec string edit distance	13%
Alignment label matches	88%
Alignment label matches (only changed arcs)	65%
Same verbs annotated	85%
Same verb role types labeled	70%

Table 2: Some pairwise inter-annotator agreement statistics for manual reconstructions of Fisher data.

The finished annotation product yielded several interesting observations. Pairwise comparisons between any two reconstructions of the same string match exactly just over than half of the time (67%), though any given word in one reconstruction appears in the other reconstruction almost 95% of the time. The average string edit distance between any pair of corresponding reconstructions is 13%, which helps confirm that edit distance between a hypothesis reconstruction and any fixed reconstruction is likely to be a weak evaluation metric. 88% of the time any pair of annotators made the same reconstruction decision (alignment labels matched). Variance was attributed to the nondeterminism of the reconstruction process and indeed to specific annotation styles of the annotators, some of whom were more likely to delete than to move words, etc.

Even when reconstructed strings aren't exact matches, we observed that the verbs labeled for their semantic roles should be approximately the same if the meaning is indeed preserved in both reconstructions. For any pair of reconstructions, the same verbs were annotated 85% of the time. Examples of when this did not happen include instances of "I guess" at the end of a sentence, which were at times considered to not contribute to the meaning of the sentence and deleted as fillers, and were at other times preserved. In other instances, annotator error was to blame: sometimes verbs were missed and their arguments left unlabeled.

5. Conclusions

The Spontaneous Speech Reconstruction corpus produced in this work includes 6,384 spontaneously spoken sentence-like units, each annotated twice for quality control and future evaluation mechanisms. This resource supplements previously existing LDC manually generated parse trees, transcripts, and edit labels for a subsection of the Fisher corpus. The additions include sentence-level reconstruction with word-level alignments with labels, to be used for future research into deep sentence cleanup for spontaneous speech, and predicate-argument labels for all grammatical sentences which, combined with the reconstruction alignments, may yield new quantifiable insights into the structure of disfluent natural speech text.

We intend to continue building and revising the SSR corpus and hope to make the data available publicly soon via http://www.clsp.jhu.edu/research/pire/ssr/. We hope that the rich data set may facilitate new research efforts in the area of reconstructing and representing the structure of spontaneously produced speech.

6. Acknowledgements

The authors would like to acknowledge the help of Petr Podveský and Petr Pajas of Charles University in building the annotation tool used in this work. Thanks also to Olga Babko-Malaya of the Linguistic Data Consortium for her help in answering questions about predicate-argument annotation and advising in the initial annotation process.

7. References

Ulrich Callmeier. 2001. Efficient parsing with large-scale unification grammars. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

Christopher Cieri, Stephanie Strassel, Mohamed Maamouri, Shudong Huang, James Fiumara, David Graff, Kevin Walker, and Mark Liberman. 2004. Linguistic resource creation and distribution for EARS. In *Rich Transcription Fall Workshop*, number 9.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

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

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chiacgo Press and CSLI Publications, Chicago and Stanford.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association of Computational Linguistics (HLT/NAACL)*, Boston, MA.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley.

Stephanie Strassel and Christopher R. Walker. 2004. Linguistic resources for metadata extraction. In *Rich Transcription Fall Workshop*, number 39.