PDTSL: AN ANNOTATED RESOURCE FOR SPEECH RECONSTRUCTION

Jan Hajič, Silvie Cinková, Marie Mikulová, Petr Pajas, Jan Ptáček, Josef Toman, Zdeňka Urešová

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské nám. 25, 11800 Prague 1, Czech Republic

ABSTRACT

We present a description of a new resource (Prague Dependency Treebank of Spoken Language) being created for English and Czech to be used for the task of speech understanding, broad natural language analysis for dialog systems and other speech-related tasks, including speech editing. The resources we have created so far contain audio and a standard transcription of spontaneous speech, but as a novel layer, we add an edited ("reconstructed") version of the spoken utterances. These edits go beyond the scope of current speech reconstruction efforts in that we allow, on top of the usual deletions of speech artifacts, fillers, etc. also for word modifications, insertions and word order changes. We have used both monologue and dialogue recordings in English and Czech to verify the feasibility of such transcription. We have also assessed the quality of the resulting annotation since the relative freedom of the editing raises an issue of what a "correct" annotation is.

Index Terms— speech recognition, speech reconstruction, spoken data resources, annotation

1. INTRODUCTION

Current automatic speech recognition (ASR) paradigm uses the truthfulness of the automatic transcription as the main objective which also serves as the main basis for evaluation of ASR systems. However, with the increasing accuracy of large vocabulary speech recognition systems, the gap widens between what is actually necessary for further broad natural language analysis (in systems that cannot rely on fixed grammars) and the exact transcription. While linguists might like the exact placement and recognition of all words, pauses and non-speech events that some systems are capable of outputting, for subsequent processing in speech understanding systems (such as broad-coverage dialog systems) such a precision is not necessary. On the other hand, current language analysis tools taken over from the written text domain assume, despite their robustness based on statistical training, that the input text is more or less grammatical, or at least close to what they have been trained on - namely, written-text (plain or manually annotated) resources. Taggers, parsers, word sense disambiguators, noun phrase chunkers, name entity recognizers and machine translation systems are all trained on resources that come from the written text domain, and their drop in accuracy is significant if presented with speech recognizer output (even with retraining on domain- and style-specific data). It would not only be costly to re-annotate spoken data for re-training these tools on spoken data, but often it would mean to redefine even the annotation guidelines, since spoken language is, to say the least, much less "regular" than written language.

The research described herein has been supported by the grant GA405/06/0589 of the Grant Agency of the Czech Republic, projects No. MSM0021620838 and ME838 of the Ministry of Education of the Czech Republic, by the Charles University Grant Agency grants GAUK 52408 and 22908, the NSF PIRE grant No. 0530118 and by the EU grant IST-FP6-034434.

2. THE DATA

2.1. The English Corpus

The English corpus has been recorded under semi-controlled conditions and currently contains about 70 hours of audio collected at Napier University [1]. It mainly consists of short-turn dialogs with some longer passages of single-speaker-produced speech. There is a single topic of these dialogs, namely, a moderator and an invited person talk over photos from the person's archive. It has originally been created for the "Companions" EU project [2].

Out of the 70 hours recorded, about 20 hours have been transcribed using the standard transcription tool "Transcriber" [3]. This is our English source material; from that, we have so far manually annotated (reconstructed) 105,000 tokens in 8,500 sentences..

2.2. The Czech Corpus

For Czech, we have used the Czech portion of the Malach project [4] corpus. The Czech Malach corpus consists of lightly moderated dialogs (interviews or "testimonies") with

Holocaust survivors, originally recorded for the Shoa memory project by the Shoa Visual History Foundation, now hosted by the University of Southern California. The dialogs usually start with shorter turns but continue as longer monologues by the survivors, often showing emotion, disfluencies caused by recollecting interviewee's distant memories, etc.

From the 576 interviews originally recorded, 80 hours have been manually transcribed using (also, as in the English dialogs described above) the standard Transcriber tool. This is our Czech source material at the moment; from that, we have manually annotated (reconstructed) 268,000 tokens, most of it double-annotated. We have already started the annotation of the Companions topical dialogs (recorded in Czech) for direct, domain-identical comparison with English.

3. THE ANNOTATION

3.1. Previous Work

The fact that spontaneous speech is "ungrammatical", full of disfluencies, false starts, repeats, fillers etc. is obvious to anyone who has listened to any speech recording (there are exceptions, which we found even in our spontaneously recoded data, but they are extremely rare). For more than a decade, these phenomena have caused problems for any subsequent processing other than pure word-error-rate-based automatic speech recognition competitions - be it for dialog systems [5], [6] or for extracting meaningful text pieces [7]. Most recently, [8] uses a rich and extensive annotation scheme over the Fisher corpus and provides experimental results in identifying disfluencies.

All these projects aim at identifying and labeling segments of the original audio (and transcription) for the chosen disfluencies. Based on the labelings, they can then be "corrected" using (almost exclusively) deletions from the transcription. Insertion of punctuation has also been attempted (e.g., before parsing is attempted) but it usually assumes that disfluencies are already being solved, or simply ignores them.

However, this style of disfluency identification (and correction) cannot, in general, arrive at grammatical, fluent text, even though the previous results are often impressive and certainly make the text much more understandable for the human reader as well as for subsequent automatic analysis.

3.2. Annotation Principles and Guidelines

Our aim was thus to prepare data where the manually prepared fluent and grammatical text will be linked to both the automatically and manually transcribed audio, to be later used for various machine learning experiments and thus lead to tools for full automatic speech "reconstruction" - for more details, see Sect. 7. This paper describes the first step, namely the reconstructed text annotation and linkage; for future plans, see Sect. 8.

In order to achieve a fully fluent text we have to allow for any changes that this goal requires, most notably (and in addition to the usual approach as shown in the previous work described above in Sect. 3.1), we do allow for token (word) changes (including proper digitization of spoken numbers and numerical expressions), word reordering, and word and punctuation insertions.

On the other hand, the annotation guidelines do not call for extensive labeling of all the changes and differences; most labels can easily be derived from the links (or their absence) between our manual reconstruction annotation and the standard transcription which we are getting together with the source data as described in Sect. 2. Other labels that are more syntactically and semantically oriented, such as those used in the previous projects by e.g. [7] and [8], are not used at this stage, since they will be obtained by back linking future syntactic and semantic annotation (see Sect. 8).

The annotators are thus simulating the work of magazine editors when preparing recorded interviews to appear in printed form. In fact, this idea is almost all that the annotators are taught before they can start annotating; specifically, there are two basic annotation principles they have to follow:

- the output must preserve the meaning of the source as much as possible while being grammatically correct, and properly segmented, fluent and stylistically appropriate English or Czech as one would expect it to appear in a printed interview, and
- the amount of changes to the original (true) audio transcription should be minimal.

We thus do not give the annotators any source-material-based rules, such as that they should tie their annotation to prosody patterns, or follow pauses etc.

However, since synchronization can no longer be used (or only with heavy difficulties) due to possible word order changes, word and punctuation insertions etc., the annotators are required to correctly link the reconstructed text tokens to the original transcription (which is, of course, then linked implicitly by using the synchronization marks to both the automatically recognized audio (if it exists) and to the audio itself). Even though the rules are relatively simple, certain conventions had to be introduced:

- source deletions: not linked
- word and punctuation insertions: not linked
- word substitution changes: linked to the source token that is the most similar; if indistinguishable, to the last one present in source (for repeated tokens)
- word orthographical changes only or no change (identity between source and annotation): links to the source token (rightmost if repetitions in source).

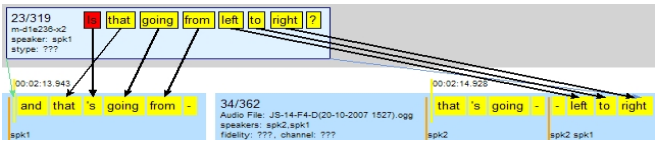

Fig. 1. Example of speech reconstructed annotation (as seen in the MEd annotation tool)

By orthographical change we mean a change in capitalization only, otherwise the change (such as a correction of a colloquial form if deemed necessary) is labeled as "substantial". Word order changes are not labeled since they are deterministically extractable from the (crossing) links. Such "implicit" labeling not marked by the annotators (word order changes, insertions, deletions) will be introduced in the public release of the data for their easier use.

Some other simple conventions are introduced for resegmentation and resynchronization of the source into meaningful sentences within turns, for marking overlapped speech etc.

The same general annotation guidelines are applied for both Czech and English, and they are general enough to be applied to other languages as well (respecting, of course, individual lanuages' standard orthography rules, format of numbers, etc.). An example of an annotated segment with several important changes can be seen in Fig. 1.

4. THE MARKUP AND DATA FORMAT

The markup used in the reconstruction annotation is specified using the language independent Prague Markup Language (PML) [9], which is an XML-based metalanguage for language and speech analysis annotation used in the Prague Dependency Treebank (PDT) v. 2.0 [10]. The PML adds semantics to linear and tree-based schemes of natural language annotation and allows for interlinked hierarchical layers of standoff annotation. There are four layers in the PDT: the token layer, the morphological layer, the surface syntactic layer, and the syntactico-semantic ("tectogrammatical") layer. For the speech reconstruction, we have added the audio as an external base layer and we have extended the token layer to contain the manual speech transcription, and the morphological layer to contain the speech-reconstructed annotation (see Fig. 2). For the automatic speech recognition output of the audio (if available) a separate, simplified token layer has been introduced, which can be interlinked with the manual transcription using the synchronization points.

5. TOOLS FOR MANUAL ANNOTATION

The editor MEd [11] is the main annotation tool that is being used for the speech reconstruction annotation. It is a second generation tool (modeled after the first such tool, used by [8] in their annotation efforts). MEd can handle PML directly,

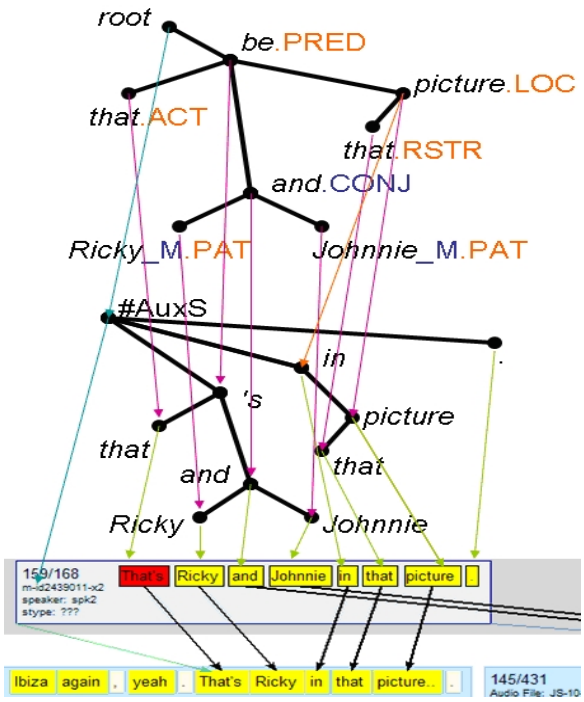

Fig. 2. Layers of annotation (from the top: semantics, syntax, reconstructed speech, transcription)

and can work with all of the audio, ASR transcription, manual transcription and the speech-reconstruction annotation at the same time. Essentially, its functionality extends that of the Transcriber tools in that it allows for arbitrary editing of the transcription and for linking the edited text back to the transcription and the audio.

6. ANNOTATION QUALITY

Speech annotation, even a simple transcription, is not a simple task and errors of various kinds can be introduced easily. Putting aside performance errors that can be handled in known ways, we will discuss briefly the types of interannotator disagreement that are specific to the speech reconstruction annotation style described here.

In general, the less rigid the annotation guidelines are, the more disagreement among annotators can be expected. In our case, the guidelines are in fact very vague in the main points (cf. Sect. 3.2), namely the "meaning preservation" principle and the "minimal changes" principle. On the other hand, punctuation, orthography and grammaticality and some of the lining rules can be checked to varying extent using text-annotation-based methods. Moreover, the specification of the data in PML in cooperation with MEd does not allow for the typical "markup errors" at all.

Given the above, we have decided to keep multiple annotations for each source dataset (for Czech, we have almost completed double annotation. For both languages, we aim at

triple annotation). Using these multiple annotations, we can check for the "mechanical" and standard language errors, but we will NOT unify the individual annotation streams beyond that. We believe that it will, in the long run, lead to more possibilities of training and evaluation of any tools that might be developed using such data, in a similar vein to the way multiple reference translations are used for automatic machine translation evaluation, except in this case the multiple annotated streams will be available for the statistical learning phase as well.

7. INTENDED USE

In the first release of the data, we will complement the speech-reconstruction annotation with automatically added syntactic and semantic annotation, using currently available state-of-the-art tools (syntactic and semantic parsers) for both Czech and English. The main purpose is to allow for more fine-grained experiments with the undoubtedly more difficult speech reconstruction task using the data as described above (due to the word order and other "difficult" changes), using the rich syntactic and semantic features from the parsers mapped back to the audio and/or ASR transcription layers.

This is possible thanks to the PML links through which any semantic, syntactic and morphological feature can be traced all the way back even to the audio signal (of course, the underlying speech segment does not have to be continuous - one can get multiple segments being linked to a single token or feature).

We are currently in the process of using some of the features in a simple machine-translation-like speech reconstruction tool, which will form a baseline for future experiments.

8. CONCLUSIONS AND FUTURE WORK

We have presented a description of a resource for a truly edited, "reconstructed" annotation of speech input. This resource is now available for download[1] for experiments.

In the future, we will make a full-fledged LDC release containing the additional syntactic and semantic annotation. Eventually, this annotation will also be manually corrected, to allow for even more interesting experiments with syntax and semantics extraction from spontaneous speech.

9. REFERENCES

[1] J. Bradley, O. Mival, and D. Benyon, "A Novel Architecture for Designing by Wizard of Oz," in *Proceedings of CREATE08*, 2008, pp. 1–4.

[2] Proposal/Contract no.: IST-034434, European Commission (http://www.companions-project.org): Yorick Wilks (PI), "Companions: Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet," 2006–2010.

[3] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication special issue on Speech Annotation and Corpus Tools*, vol. 33, no. 1–2, pp. 5–22, 2001.

[4] William Byrne, David Doermann, Martin Franz, Samuel Gustman, Jan Hajic, Douglas Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu, "Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 420–435, July 2004.

[5] P. Heeman and James Allen, "Tagging Speech Repairs," in *ARPA Workshop on Human Language Technology*, Princeton, NJ, 1994, pp. 187–192.

[6] P. Heeman and James Allen, "Detecting and Correcting Speech Repairs," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, Las Cruces, June, 1994, pp. 295–302.

[7] J. Kolar, J. Svec, S. Strassel, C. Walker, D. Kozlikova, and J. Psutka, "Czech Spontaneous Speech Corpus with Structural Metadata," in *Proceedings of the 9th European Conference on Speech Communication and Technology, INTERSPEECH 2005*, Lisboa, Portugal, 2005, pp. 1165–1168.

[8] Erin Fitzgerald and Frederick Jelinek, "Linguistic resources for reconstructing spontaneous speech text," in *LREC Proceedings*, Marrakesh, Morocco, 2008, ELRA, pp. 1–8.

[9] Petr Pajas and Jan Stepanek, "XML-Based Representation of Multi-Layered Annotation in the PDT 2.0," in *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, 2006, pp. 40–47.

[10] Jan Hajic, Jarmila Panevova, Eva Hajicova, Petr Sgall, Petr Pajas, Jan Stepanek, Jiri Havelka, Marie Mikulova, Zdenek Zabokrtsky, and Magda Sevcikova-Razimova (Linguistic Data Consortium, Philadelphia, PA, USA), "The Prague Dependency Treebank 2.0, Cat. No. LDC2006T01," 2006.

[11] Petr Pajas and David Marecek, "MEd - an editor of interlinked multi-layered linearly-structured linguistic annotations (http://ufal.mff.cuni.cz/~pajas/med)," 2007.

[1] http://ufal.mff.cuni.cz/~hajic/pdtsl