Semi-automatically Alignment of Predicates between Speech and OntoNotes Data

Niraj Shrestha, Marie-Francine Moens
Department of Computer Science,
KU Leuven, Belgium
{niraj.shrestha, Marie-Francine.Moens}@cs.kuleuven.be

Abstract

Speech data currently receives a growing attention and is an important source of information. We still lack suitable corpora of transcribed speech annotated with semantic roles that can be used for semantic role labeling (SRL), which is not the case for written data. Semantic role labeling in speech data is a challenging and complex task due to the lack of sentence boundaries and the many transcription errors such as insertion, deletion and misspellings of words. In written data, SRL evaluation is performed at the sentence level, but in speech data sentence boundaries identification is still a bottleneck which makes evaluation more complex. In this work, we semi-automatically align the predicates found in transcribed speech obtained with an automatic speech recognizer (ASR) with the predicates found in the corresponding written documents of the OntoNotes corpus and manually align the semantic roles of these predicates thus obtaining annotated semantic frames in the speech data. This data can serve as gold standard alignments for future research in semantic role labeling of speech data.

Keywords: Speech data, Semantic role labeling, Semantics, Automatic speech recognition

1. Introduction

Semantic role labeling (SRL) is a process of predicting the predicate-argument structures in language utterances by identifying predicates and their related semantic roles. SRL reveals more information about the content than a syntactic analysis in the field of natural language processing (NLP) in order to better understand "who" did "what" to "whom", and "how", "when" and "where".

SRL has many key applications in NLP, such as question answering, machine translation, and dialogue systems. Many effective SRL systems have been developed to work with written data. However, when applying popular SRL systems such as ASSERT (Pradhan et al., 2005), Lund SRL (Johansson and Nugues, 2008), SWIRL (Surdeanu and Turmo, 2005), and Illinois SRL (Punyakanok et al., 2008) on transcribed speech, which was generated by an automatic speech recognizer (ASR), many errors are made due to the specific nature of the ASR transcribed data.

SRL on written data performs well due to the availability of annotated corpora like PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998) etc., which help to train the SRL system and also the written data is clean and well-formed. Most of the SRL systems on written data are evaluated at the sentence level. On the other hand, ASR data is noisy and not well-formed, it does not contain sentence boundaries and it exhibits many errors like insertion, deletion and misspelling of words. Because of the lack of sentence boundaries in speech data and the problem of transcribed speech, it is very hard to align speech data and OntoNotes data on the sentence level. Keeping these complexities in mind, we align predicates and their semantic roles between speech data and their corresponding written texts of the OntoNotes corpus, more specifically the OntoNotes release 3.0 dataset which covers English broadcast and conversation news, so that this resource can be used in future research work.

In this way, we have created predicate argument structures in the speech data which function at the semantic frame level rather than at the sentence level. The annotated speech meta-data can be downloaded from here[1] for research purposes. The corresponding speech data can be obtained from (Favre et al., 2010).

2. Building the Resource

2.1. OntoNotes Data

The OntoNotes data is in CoNLL-like format. Each line represents a token and an empty line represents the end of a sentence. A line has many columns that represent a linguistic feature like lemma form, part-of-speech (POS), parse information, predicate argument structure etc. including the token itself and the token's id. The number of columns for predicate argument annotation is variable i.e. one per each predicate and the annotation starts from the twelfth column. If there are no predicates annotated in a sentence then that column is labeled with "*".

```
0 A * * *
1 much (ARGM-MNR* * *
2 better *) * *
3 looking (V*) * *
4 News (ARG0* * *
5 Night *) * *
6 I * (ARG0*) *
7 might * (ARGM-MOD*) *
8 add * (V*) *
9 as * * *
10 Paula * * (ARG1*
11 Zahn * * *)
12 sits * * (V*)
13 in * * *
14 for * * (ARG2*
```

[1] https://people.cs.kuleuven.be/∼niraj.shrestha/speechData

```
A [ARGM-MNR much  better] [TARGET looking] [ARG0 News  Night] I might
add as Paula Zahn sits in for Anderson and Aaron .

A much better looking News Night [ARG0 I] [ARGM-MOD might] [TARGET
add] as Paula Zahn sits in for Anderson and Aaron .

A much better looking News Night I might add as [ARG1 Paula  Zahn]
[TARGET sits] in [ARG2 for Anderson and  Aaron] .
```

Figure 1: An example sentence from OntoNotes with its predicate argument structures for three predicates viz. *looking*, *add* and *sits*.

Table 1: Statistics of predicate alignment between the gold standard and the speech data.

	Frequency
# Predicates alignment between gold standard and speech data	
# Predicates occurs only once	17220
# Predicates aligned left right context	24278
# Predicates aligned manually	3541
# Total alignments	**45039**
# Predicates not aligned or not found	**7142**
# Predicates in gold standard data	**52181**

```
15 Anderson * * *
16 and * * *
17 Aaron * * *)
18 . * * *
```

In the above example we only show an example predicate argument structure, which has three predicates viz. *looking*, *add* and *sits* and these predicates are represented by *(V*)* . We present the above annotation format in a more readable format as shown in figure 1. Here, each argument is enclosed in square brackets *[ ]* containing its argument name, for example *[ARGM-MNR much better]*, and the predicate is represented by the word *"TARGET"* enclosed in square brackets.

When an SRL system is applied on written data, it is evaluated at the sentence level, that means there is a one to one alignment between the ground truth written sentence and the written sentence that is annotated by the SRL system, which makes evaluation easy. But this is not the case in ASR data, where there are no sentence boundaries and identifying sentence boundaries in such data is still a bottleneck. Figure 3 shows a ASR snippet split into sentences using an automatic sentence boundary detection provided by (Favre et al., 2010) and figure 2 shows the corresponding written sentences from the OntoNotes corpus. From these two figures, it is observed that the ASR data are split at the wrong position or two sentences are merged into one. This makes the evaluation with ground truth annotation very difficult and we need the alignments of predicates and their semantic roles between the ground truth data and the ASR data.

2.2. Data Set

We use the OntoNotes release 3.0 dataset which covers English broadcast and conversation news (Hovy et al., 2006). We align the predicates between the speech data and the OntoNotes data for the subset. The subset data contains {/bc/cnn, /bc/msnbc, /bn/abc, /bn/cnn, /bn/mnb, /bn/nbc, /bn/pri, /bn/voa} of OntoNotes release 3.0 as used by (Favre et al., 2010) since we have speech transcriptions for that subset only. This subset constitutes our gold standard dataset.

Table 2 shows the statistics of the predicates in the gold standard dataset. We have all together 722 files, in which there are 52181 predicates in total. The maximum number of predicates in a file is 2105, while the minimum is 2. So, on average there are 72 predicates per file.

3. Alignment Procedure

The alignment task of predicates and their semantic roles is challenging because the transcribed speech does not contain sentence boundaries. First, we align the predicates between the gold standard data and the speech data after that we align their corresponding semantic roles.

3.1. Predicate Alignment

Before aligning the predicates between two corpora, we assign a token id for each token in both corpora so that the predicate also has token id in both corpora. For each pred-

Table 2: Predicate statistics in the gold standard dataset.

	Frequency
# Total number of files	722
# Total number of predicates	52181
# Maximum number of predicates in a file	2105
# Minimum number of predicates in a file	2
# Average number of predicates in a file	72

A much better looking News Night I might add as Paula Zahn sits in for Anderson and Aaron . \\
They 're both off - \\
Look at that . \\
Is that a replacement ? \\
Paula Paula . \\
Thank you for your faith Larry and thank you for your graciousness . \\
And we 're going to get started here . \\
Good evening everybody . \\

Figure 2: Sentences from OntoNotes data (" \\" represents the end of line).

a much better looking newsnight i might add \\
as powerless on sits in for anderson \\
and they're and they're both off of that that is not a replacement \\
colin powell \\
thank you for your faith larry \\
and thank you for your graciousness will give and we're going to get started here good evening \\

Figure 3: Automatic segmented sentences from ASR data (" \\" represents the end of line).

icate from the gold standard, we look for the matching token in the speech data within its three left and three right context words. If a predicate in the gold standard occurs once and matches with a token in the speech data, they are aligned using the one-to-one principle without taking into account the predicate's left and right context. The accuracy of one-to-one predicate alignment is 100% measured in a sample of 100 predicates. If a predicate appears more than once in either corpus, the predicate from the speech data is aligned with a predicate in the gold standard, if their words in the left and right context matches. If their left and right contexts do not match, we have aligned the predicate manually. In Figure 4, *example a* shows the one-to-one predicate alignment between ground truth and the ASR data with its three words left and right context matching. But this is not always the case where predicate's left and right context will match because of speech transcription errors as shown in figure 5, example 1. In such cases we have manually aligned the predicates. Sometimes, we are not able to align the predicate from ground truth to ASR data due to ASR errors as shown in figure 5 example 2, where the predicate *"went"* is wrongly transcribed to *"when"*, so we miss such predicates and its semantic roles.

Table 3: An example of predicate alignment between the gold standard (GT) and the speech dataset (SYS).

SYS's predicate	**SYS_tokenID**	**GT's predicate**	**GT_tokenID**
attended	6126	attended	6754
responded	6138	responded	6766
care	5597	care	6104
care	6630	care	7349
care	7145	care	7916
seems	5881	seems	6450
seems	6674	seems	7395
seems	5287	seems	5758

Table 3 shows a snippet of predicate alignments between the gold standard and the speech data. Predicates *"attended"* and *"responded"* are matched in one to one fashion while predicates *"care"* and *"seems"* are aligned based on their left and right three words context.

It is shown in table 1 that there are 52181 predicates contained in the gold standard. We are able to align 45039 predicates from the speech data with the gold standard data, out of which 17220 predicates have been aligned in the one-to-one fashion, 24278 predicates have been aligned relying on the left and right context matching. The accuracy of left and right context matching alignment is 90% measured in a sample data of 100 predicates. We have manually aligned the remaining 3541 predicates. This constitutes our predicate alignment between the speech data and the gold standard. We have also noticed that 7142 predicates of the gold standard data are not aligned with speech data, which is due to speech recognition errors and especially missing words in the transcribed speech.

3.2. Semantic Role Alignment

Once we have aligned the predicates in the two corpora, we have aligned manually their semantic roles. For each semantic role of a predicate, we look for its exact token matching in the speech data and if the tokens are found then we assigned the semantic role label from the OntoNotes data to the matched tokens of the speech data. Currently, we only aligned exact matches of the tokens that constitute the semantic roles. Figure 4, *example b* shows this scenario, where semantic roles *ARG0* is assigned to exact tokens matching to *"Queen Elizabeth and Prime Minister Tony Blair"* and similarly for semantic roles *ARGM-ADV* to *"as did relatives to those who died and emergency workers who responded to the terrorist attacks"*. If there are missing tokens in the speech data due to transcription errors then

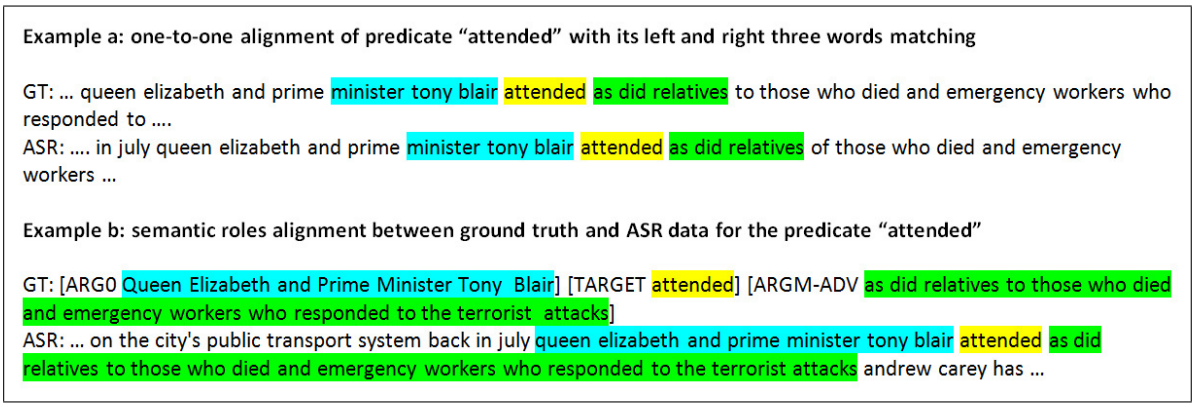

Figure 4: Examples showing the alignment between predicates and their semantic roles between ground truth and ASR data, where *GT* is ground truth data and *ASR* is speech transcript data.

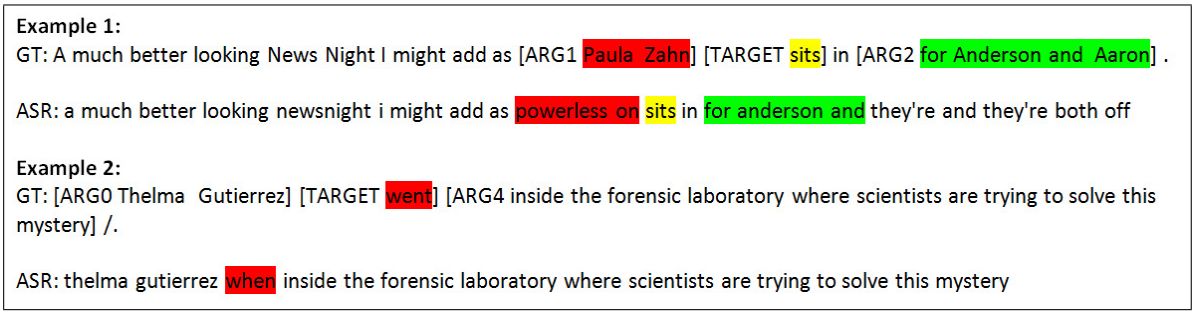

Figure 5: Example sentences showing the mismatch alignment between predicates and their arguments, where *GT* is ground truth data and *ASR* is speech transcript data.

we assume that a semantic role is missing and do not realize any alignment as shown in figure 5 *example 1*. Here the tokens *"Paula Zahn"* is wrongly transcribed to *"powerless on"* and the token *"aaron"* from ground truth is missing in the ASR data. So we are not able to align *ARG1* and *ARG2* for the predicate *"sits"*.

4. Conclusion

In this work, we have realized a predicate-argument alignment between the gold standard data and the speech data for a subset of the OntoNotes corpus. The main objective of this work was to create a dataset that copes with the lack of sentence boundaries in speech data and that can be used to evaluate the semantic role labeler at a frame level. We hope that this data set might be very useful for future NLP research. Regarding the missing predicates from the speech data, we can not do anything but as a future work we could refine this dataset by adding partial matching of semantic roles between the gold standard data and speech data. This task might be difficult as it is hard to decide the minimum amount of matching of a semantic role in the speech data to be valid for semantic role annotation. So we leave this as an open question to be discussed in future work.

5. Acknowledgements

This work is financially supported by ACquiring CrUcial Medical information Using LAnguage TEchnology (ACCUMULATE) project (IWT-SBO 150056).

6. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Favre, B., Bohnet, B., and Hakkani-Tür, D. (2010). Evaluation of semantic role labeling and dependency parsing of automatic speech recognition output. In *Acoustics Speech and Signal Processing (ICASSP), Proceedings of the 2010 IEEE International Conference on*, pages 5342–5345, March.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johansson, R. and Nugues, P. (2008). Dependency-based semantic role labeling of propbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 69–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March.

Pradhan, S., Hacioglu, K., Ward, W., Martin, J. H., and Jurafsky, D. (2005). Semantic role chunking combining complementary syntactic views. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 217–220, Stroudsburg, PA, USA.

Punyakanok, V., Roth, D., and Yih, W. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Comput. Linguist.*, 34(2):257–287, June.

Surdeanu, M. and Turmo, J. (2005). Semantic role labeling using complete syntactic analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 221–224, Stroudsburg, PA, USA. Association for Computational Linguistics.