Adapting dependency parsing to spontaneous speech for open domain spoken language understanding

Frederic Bechet, Alexis Nasr, Benoit Favre

Aix Marseille Universite, CNRS-LIF

Abstract

Parsing human-human conversations consists in automatically enriching text transcription with semantic structure information. We use in this paper a FrameNet-based approach to semantics that, without needing a full semantic parse of a message, goes further than a simple flat translation of a message into basic concepts. FrameNet-based semantic parsing may follow a syntactic parsing step, however spoken conversations in customer service telephone call centers present very specific characteristics such as non-canonical language, noisy messages (disfluencies, repetitions, truncated words or automatic speech transcription errors) and the presence of superfluous information. For syntactic parsing the traditional view based on context-free grammars is not suitable for processing non-canonical text. New approaches to parsing based on dependency structures and discriminative machine learning techniques are more adapted to process spontaneous speech for two main reasons: (a) they need less training data and (b) the annotation with syntactic dependencies of conversation transcripts is simpler than with syntactic constituents. Another advantage is that partial annotation can be performed. This paper presents the adaptation of a syntactic dependency parser to process very spontaneous speech recorded in a call-centre environment. This parser is used in order to produce FrameNet candidates for characterizing conversations between an operator and a caller.

Index Terms: dependency parsing, FrameNet, spoken language understanding, spontaneous speech.

1. Introduction

Parsing human-human conversations consists in automatically enriching text transcription with semantic structure information. Such information includes sentence boundaries, syntactic and semantic parse of each sentence, para-semantic traits related to several paralinguistic dimensions (emotion, polarity, behavioural patterns). Spoken conversations in customer service telephone call centers present specific characteristics such as:

- non-canonical language: spontaneous spoken conversations represent different levels of language than the "canonical" one used in written text such as newspaper articles;
- "noisy messages": spoken conversation transcriptions may contain disfluencies, repetitions, truncated words or automatic speech transcription errors;
- superfluous information: redundancy and digression make conversation messages prone to contain superfluous information that needs to be discarded;
- conversation transcripts are not self-sufficient: for spoken messages, even with a perfect transcription, suprasegmental information (prosody, voice quality) has to be added to the transcription in order to convey speakers intention (sentiment, behaviour, polarity).

Semantic parsing is the process of producing semantic interpretations from words and other linguistic events that are automatically detected in a text conversation or a speech signal. Many semantic models have been proposed, ranging from formal models encoding "deep" semantic structures to shallow ones considering only the main topic of a document and its main concepts or entities. We will use in this study a FrameNet-based approach to semantics that, without needing a full semantic parse of a message, goes further than a simple flat translation of a message into basic concepts: FrameNet-based semantic parsers detect in a sentence the expression of frames and their roles. Because frames and roles abstract away from syntactic and lexical variation, FrameNet semantic analysis gives enhanced access to the meaning of texts: (of the kind "who does what, and how where and when ?").

FrameNet-based semantic parsing is often based on a syntactic parsing step. However, for processing noncanonical text such as automatic speech transcripts, the traditional view of parsing based on context-free grammars is not suitable: due to ungrammatical structures in this kind of text, writing a generative grammar and annotating transcripts with that grammar is difficult. New approaches to parsing based on dependency structures and discriminative machine learning techniques [1] are more appropriate for two main reasons: (a) they need less training data and (b) the annotation with syntactic dependencies of conversation transcripts is simpler than with syntactic constituents.

Using dependency parsing for speech processing has been proposed in previous studies ([2, 3]), however the problem of the *adaptation* of a dependency parser to the specificities of speech transcripts, manual or automatic, of spontaneous real-world speech remains an open problem.

This paper describes the adaptation process of a dependency parser to spontaneous speech in order to perform open domain Spoken Language Understanding thanks to a FrameNet approach. We will present why it is crucial to adapt parsers that are originally trained on written text to the specificities of spontaneous speech on manual transcriptions containing disfluencies, and discuss the usefuleness of this approach to perform open-domain SLU on ASR transcriptions even with a high WER. All the experiments have been carried on the RATP-DECODA corpus containing recordings of conversations in the Paris public transport authority call-centre.

2. Related work

Many methods have been proposed for limited domain SLU, following early works on the ATIS corpus (see [4] for a review of SLU methods and models). Regardless of the paradigm chosen for performing SLU (parsing, classification, sequence la-

belling), the domain-ontology concepts and relations are always directly predicted from the ASR word transcriptions, sometimes with features coming from a linguistic analysis based on generic syntactic or semantic models. For open-domain SLU, it is necessary to choose an abstract level of representation that can be applied to a large range of domains and applications, therefore syntactic and semantic models developed in the Natural Language Processing community for processing text input are good candidates.

As presented in the introduction, we choose a FrameNet approach to semantic in this paper. FrameNet parsing is traditionnaly decomposed into the following subtasks (whether applied sequentially or not):

- trigger identification: find the words that express frames. For instance in "*she declared to her friend that she was going out*". The target word "*declared*" is identified.
- trigger classification: assign the relevant frame in context (assign the frame STATEMENT to the trigger "*declared*")
- role filler identification: find/segment the expressions that may fill a frame role ("*she*", "*to her friend*" and "*that she was going out*" should be identified as potential role fillers
- role filler classification: assign the roles to the role fillers candidates ("*she*", "*to her friend*" and "*that she was going out*" play respectively the Speaker, Addressee and Message roles, defined for the frame STATEMENT

The last two subtasks are generally referred to as " *semantic role labeling* " (SRL), though this term is more general and includes SRL with other roles than that of FrameNet, in particular PropBank roles. [5] presented the first study on role filler classification: they proposed a probabilistic classifier that, given an English sentence, a lexical trigger within that sentence and the (gold) corresponding frame, assigns FrameNet roles to syntactic phrases within the sentence. This seminal work was followed by a large number of studies, with variants using other kinds of classifiers such as maximum entropy [6] or SVM [7]. /storage/raid2

Recently, in [8], a FrameNet parser was used to process spontaneous speech for the development of a Spoken Dialog System. However, in contrast with our study, no adaptation to the specificities of spontaneous speech was perfomed on the linguistic models of the parser: the authors used the *SEMAFOR* [9] parser trained on written text. The main claim of this paper is to highlight the need for such an adaptation process when dealing with real-world spontaneous speech because of two main issues:

1. firstly spontaneous speech transcriptions are often difficult to parse using models developed for written text due to the specificities of spontaneous speech syntax (agrammaticality, disfluences such as repairs, false starts or repetitions);
2. secondly, transcriptions obtained through an Automatic Speech Recognition (ASR) process contain errors, the amount of errors increasing with the level of spontaneity in speech.

The first issue can be partially tackled by using new approaches to parsing. Syntactic parsing aims to uncover the word relationships (e.g. word order, constituents) within a sentence and support the semantic layer of the language-processing pipeline. Parsing is traditionally tightly connected to rewriting grammars, usually context free grammars, used together with a disambiguation model. Many current state-of-the-art text parsers are built on this model, such as [10]. Shallow syntactic processes, including part-of-speech and syntactic chunk tagging, are usually performed in the first stage. This traditional view of parsing based on context-free grammars is not suitable for processing non-canonical text such as automatic speech transcripts: due to ungrammatical structures in this kind of text, writing a generative grammar and annotating transcripts with that grammar is difficult.

New approaches to parsing based on dependency structures and discriminative machine learning techniques [11] are much easier to adapt to non-canonical text for two main reasons: they need less training data and the annotation with syntactic dependencies of spoken transcripts is simpler than with syntactic constituents. Other advantages are the fact that partial annotation can be performed [2] and the parses generated are much closer to meaning than constituent trees, which eases semantic interpretation.

For the second issue of ASR errors and syntactic parsing, most of the work have addressed this problem from a different point of view: using syntactic features during ASR to help reducing Word Error Rate (WER). This can be done by directly integrating parsing and ASR language models [12] or keeping them as separate processes through a reranking approach using both ASR and parsing features [3, 13]. The improvement in ASR transcriptions obtained by adding syntactic features to the models is often rather small, however the structure and the relations between words obtained through parsing can be of great interest for the SLU processes, even without a significant decrease of WER.

3. A corpus of call-centre conversations

The RATP-DECODA[1] corpus consists of 1514 conversations over the phone recorded at the Paris public transport call center over a period of two days [14]. The calls are recorded with independent channels for the caller and the agent, totaling over 74 hours of French-language speech. While conversations last 3 minutes on average, about a third is less than one minute, 12% are longer than 5 minutes and the longest are over ten minutes. Calls usually involve only two speakers but there can be more speakers when an agent calls another service while putting the customer on wait.

Each conversation is anonymized, segmented, transcribed, annotated with dislfuencies, POS tags and syntactic dependencies, topics and summaries. The call center dispenses information and customer services, and the two-day recording period covers a large range of situations such as asking for schedules, directions, fares, lost objects or administrative inqueries.

In the RATP-DECODA corpus, annotated disfluences consist in repetitions, discourse markers such as hesitations, and false starts. Discourse markers are the most frequent form of disfluency, occuring in 28.2% of speech segments, repetitions occur in 8.0% of segments and false starts, the least frequent, are represented in 1.1% of segments.

The DECODA corpus has been split in two subsets respectively called DEC-TRAIN (93,561 turns) and DEC-TEST (3,639 turns) that will be used for training and evaluating sys-

[1]The RATP-DECODA corpus is available for research at the Ortolang SLDR data repository: `http://sldr.org/sldr000847/fr`

tems in section 4.

4. Parsing speech

4.1. Dependency parsing

The tagger and syntactic parser we use in this study come from the MACAON tool suite [15]. The POS-tagger is based on a linear-chain CRF as implemented in the CRFsuite library [16]. The syntactic parser is a first-order graph-based dependency parser[2] trained using the discriminative perceptron learning algorithm with parameter averaging [11]. It uses the same first-order features as [17]. Compared to transition-based parsers, graph-based parsers are particularly interesting for ASR transcriptions because they have a more even distribution of errors and are less prone to error propagation [1]. This can be explained by the fact that transition-based parsers typically use a greedy inference algorithm with rich features, whereas graph-based parsers typically use exhaustive search algorithms with limited-scope features.

We used in this study the ORFEO tagset for POS and dependency labels. The ORFEO POS tagset is made of 17 tags. Words that are part of a disfluent expression have been assigned a POS. For example, a repetition such as: "*je je je veux*" (*I I I want*) is tagged: "CLI CLI CLI VRB". The ORFEO syntactic dependency labels tagset is restricted to 12 syntactic labels (*Subject*, *Direct Object*, *Indirect Object*, *Modifier*, ...) and a specific link (*DISFLINK*) for handling dislfuencies. The *DISFLINK* dependency is introduced in order to link disfluent words to the syntactic structure of the utterance. Disfluent words are systematically linked to the preceding word in the utterance. There is no deep linguistic reason for this, the only aim is to keep the tree structure of the syntactic representation. When a disfluent word starts an utterance, it is linked to an phony empty word that starts each sentence.

4.2. Dealing with spontaneous speech

We describe in this section two experiments for parsing real-life spontaneous speech transcriptions as can be found in the RATP-DECODA corpus. The first one consists in simply using a parser that has been trained on written material. In the second one a speech corpus has been semi automatically annotated and a parser has been trained on it. All experiments have been performed on DEC-TEST, which has been manually annotated using the ORFEO dependencies label tagset.

The first parser was trained on the training section of the French Treebank [18] (FTB-TRAIN). The FTB corpus is a collection of newspaper articles from the French journal *Le Monde*. The results are reported in Table 1.

corpus	FTB	RATP-DECODA		
train	FTB-TRAIN	FTB-TRAIN	FTB-TRAIN	DEC-TRAIN
test	FTB-TEST	DEC-TEST NODISF	DEC-TEST DISF	DEC-TEST DISF
UAS	87.92	71.01	65.78	85.90
LAS	85.54	64.28	58.28	83.86

Table 1: Parsing accuracy according to the training corpus (FTB-TRAIN or DEC-TRAIN) on the FTB-TEST and DEC-TEST corpus with and without disfluencies (for DEC-TEST)

[2]Although second order parsers usually yield better results on written data, our experiments showed that first order parsers behave better on oral data.

The first column reports parsing accuracy on the FTB test set, the others on the DEC-TEST corpus from which disfluencies have been manually removed (NODISF) or kept (DISF). Two standard metrics are used to measure the quality of the syntactic trees produced by the parser. The Unlabeled Attachment Score (UAS), which is the proportion of words in a sentence for which the right governor has been predicted by the parser and the Labeled Attachment Score (LAS), which also takes into account the label of the dependency that links a word to its governor.

Table 1 shows that a parser trained on written material behaves poorly on spontaneous speech: the LAS drops from 85.54 to 58.28. The performances of the parser on speech from which disfluencies has been removed are intermediate, with a LAS equal to 64.28. This result is nontheless artificial since the disfluencies have been manually removed from the parser input.

In order to adapt the parser to the specificities of oral French, we have parsed the DEC-TRAIN corpus with the parser described above and developed an iterative process consisting in manually correcting errors found in the automatic annotations thanks to a WEB-based interface [19]. This interface allows to write regular expressions on the POS and dependency tags and the lexical forms in order to correct the annotations on the whole RATP-DECODA corpus. Then the parser is retrained with this corrected corpus. When the error rate computed on a development set is considered acceptable, this correction process stops. The resulting corpus, although not perfect, constitutes our training corpus, obtained at a reasonibly low price compared to the whole manual annotation process of the corpus.

The result of the new parser are reported in column five of Table 1. As one can see, the accuracy of the new parser is far above the accuracy of the parser trained on the FTB even after the disfluencies have been removed. The performances of the parser can be compared to the performances of a parser for written data despite the fact that the parser has been trained on a partially manually corrected corpus.

Two reasons can explain this result. The first one is that the DECODA corpus has a quite restricted and specific vocabulary and the parser used is quite good at learning lexical affinities. The second one is that the DECODA corpus has a rather simple syntax with utterances generally restricted to simple clauses and less common ambiguities, such as prepositional attachment and coordination, than written texts.

5. Using Dependency parsing for open domain SLU

We use in this study a FrameNet model adapted to French through the ASFALDA project[3]. The current model, under construction, is made of 106 frames from 9 domains. Each frame is associated to a set of *Lexical Units* (LU) that can trigger the occurrence of a frame in a text.

The first step, in annotating a corpus with FrameNet, is to detect LUs and generate frame hypotheses for each detection. We did this process on the RATP-DECODA corpus and found 188,231 frame hypotheses from 94 different frame definitions. We decided in this study to restrict our model to the frames generated by a verbal LU. With this filtering we obtained 146,356 frame hypotheses from 78 different frames.

Table 2 presents the top-10 frames found in our corpus. As expected the top frames are related either to the transport domain (SPACE) or the communication domain (COM and COG).

[3]https://sites.google.com/site/anrasfalda

Domain	Frame	# hyp.
SPACE	Arriving	8328
COM-LANG	Request	7174
COG-POS	FR-Awareness-Certainty-Opinion	4908
CAUSE	FR-Evidence-Explaining-the-facts	4168
COM-LANG	FR-Statement-manner-noise	3892
COM-LANG	Text-creation	3809
SPACE	Path-shape	3418
COG-POS	Becoming-aware	2338
SPACE	FR-Motion	2287
SPACE	FR-Traversing	2008

Table 2: Top-10 frame hypotheses in the RATP-DECODA corpus

Each frame hypothesis does not necessarily correspond to a frame, most LUs are ambiguous and can trigger more than one frame or none, according to their context of occurrence. Annotating manually with frame labels a corpus like the RATP-DECODA corpus is very costly. However we claim in this study that by merging LU detection and dependency parsing, we can produce a first frame annotation of our corpus, at a very low cost if a dependency parser is available.

This process consists, for each verbal LU, in searching in the output of the parser for the dependencies (such as subject or object) of each selected verb. If no dependencies can be found we discard the LU. Otherwise we consider it as a frame candidate. This first annotation can be further refined by adding some semantic constraints on the possible dependent of a given LU, considering the domain of the corpus.

This process is done on the manual transcription of the spoken corpus and can be used to extract semantic patterns that can be looked for in ASR transcripts, as described in [2]. The next section presents the experiments done on the automatic frame annotation of the RATP-DECODA corpus thanks to this method.

6. Experiments

The first experiment has been conducted on the manual transcription of the TEST corpus. This corpus has been manually annotated with POS and syntactic dependencies. From this reference annotation we extract, in each dialogue, all verbs from the FrameNet LU lists with their dependencies. They correspond to the basic semantic structures that are needed to access to the frame level. For example, for the verb '*perdre*' (*to lose*) we can find the following examples in our corpus: *LOSE(I, metro-card)* in "*I have lost my my metro-card in ...*"; *LOSE(daughter, teddy-bear)* in "*she my daughter lost her teddy-bear in the ...*".

These dependency structures are the target of our evaluation: we measure how well we can detect them with an automatic parser instead of manual reference annotations. We compare in table 3 the performance of the two parsers presented in section 4, the one trained only on the FTB and the one adapted to the RATP-DECODA corpus. Average Precision and Recall in the detection of LUs with dependencies are presented in table 3. As we can see, the performance of the adapted parser have a much higher precision than the standard models.

The second experiment has been conducted on the Automatic Speech Recognition (ASR) transcription of the corpus. The RATP-DECODA corpus is a very challenging corpus from an ASR point of view, as many dialogues are recorded in very noisy conditions when users are calling the service in the streets, buses or metro stations. The average WER for the callers of the call-centre is 49.4% and 42.4% for the operators.

parser	**precision**	**recall**	**f-measure**
FTB	75.9	85.5	77.3
TRAIN	88.2	88.4	87.2

Table 3: Performance detection of semantic dependency structures on the manual transcriptions of the RATP-DECODA corpus.

condition	**precision**	**recall**	**f-measure**
dep1	47.4	66.4	51.4
dep2	57.3	80.3	62.7

Table 4: Performance detection of semantic dependency structures on the ASR transcriptions

Although the average WER is very high, not all dialogues have such poor performance. Table 4 shows the performance in terms of average Precision and Recall in the detection of LUs with dependencies in the ASR transcriptions. Two conditions are compared: **dep1** compares full predicate+dependency recovery, **dep2** accepts partial match on the dependencies. The performance are rather limited, however considering the high WER of the transcriptions, the adapted dependency models show some robustness in these difficult conditions.

7. Conclusion

We use in this paper a FrameNet approach to semantics that, without needing a full semantic parse of a message, goes further than a simple flat translation of a message into basic concepts. We show that a syntactic dependency parser can be adapted successfully to process very spontaneous spoken conversations recorded in a customer service telephone call centre. This parser is used in order to produce FrameNet candidates for characterizing conversations between an operator and a caller. The adaptation process improve significantly the parsing and the frame candidate generation performance on manual transcriptions. This method, applied to very noisy ASR transcriptions, shows also a certain level of robustness. Nevertheless, the automatic annotation of the reference transcription corpus allows the SLU process to use more domain-specific models, directly infered from the annotated corpus, for processing such high WER transcripts.

8. Acknowledgements

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n 610916 SENSEI.

9. References

[1] R. T. McDonald and J. Nivre, "Characterizing the errors of data-driven dependency parsing models." in *EMNLP-CoNLL*, 2007, pp. 122–131.

[2] F. Béchet and A. Nasr, "Robust dependency parsing for spoken language understanding of spontaneous speech," in *Proc. Interspeech*. ISCA, 2009.

[3] B. Lambert, B. Raj, and R. Singh, "Discrimintively trained dependency language modeling for conversational speech recognition," in *Proc. Interspeech*. ISCA, 2013.

[4] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[5] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Comput. Linguist.*, vol. 28, no. 3, pp. 245–288, Sep. 2002. [Online]. Available: http://dx.doi.org/10.1162/089120102760275983

[6] M. Fleischman, N. Kwon, and E. Hovy, "Maximum entropy models for framenet classification," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003, pp. 49–56.

[7] B. Coppola, A. Moschitti, and G. Riccardi, "Shallow semantic parsing for spoken language understanding," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, 2009, pp. 85–88.

[8] Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky, "Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 120–125.

[9] D. Das, D. Chen, A. F. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *Computational Linguistics*, vol. 40, no. 1, pp. 9–56, 2014.

[10] S. Petrov and D. Klein, "Learning and inference for hierarchically split pcfgs," in *Proccedings of the National Conference on Artificial Intelligence*, vol. 22, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 1663.

[11] R. McDonald, K. Crammer, and F. Pereira, "Online Large-Margin Training of Dependency Parsers," in *Association for Computational Linguistics (ACL)*, 2005.

[12] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech & Language*, vol. 14, no. 4, pp. 283–332, 2000.

[13] M. Collins, B. Roark, and M. Saraclar, "Discriminative syntactic language modeling for speech recognition," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 507–514.

[14] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot, "Decoda: a call-centre human-human spoken conversation corpus." in *LREC*, 2012, pp. 1343–1347.

[15] A. Nasr, F. Béchet, J. Rey, B. Favre, and J. Le Roux, "Macaon: An nlp tool suite for processing word lattices," *Proceedings of the ACL 2011 System Demonstration*, pp. 86–91, 2011.

[16] N. Okazaki, "Crfsuite: a fast implementation of conditional random fields (crfs)," 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/

[17] B. Bohnet, "Very high accuracy and fast dependency parsing is not a contradiction," in *Proceedings of ACL*, 2010, pp. 89–97.

[18] A. Abeillé, L. Clément, and F. Toussenel, "Building a treebank for french," in *Treebanks*, A. Abeillé, Ed. Dordrecht: Kluwer, 2003.

[19] T. Bazillon, M. Deplano, F. Bechet, A. Nasr, and B. Favre, "Syntactic annotation of spontaneous speech: application to call-center conversation data." in *LREC*, 2012, pp. 1338–1342.