Enhancing The RATP-DECODA Corpus With Linguistic Annotations For Performing A Large Range Of NLP Tasks

Carole Lailler[(1)], Anaïs Landeau[(1)], Frédéric Béchet[(2)], Yannick Estève[(1)], Paul Deléglise[(1)]

Laboratoire Informatique de l'Université du Maine (LIUM)
University of Le Mans, France
Aix Marseille Universite (AMU), CNRS-LIF, France
[(1)] firstname.lastname@lium.univ-lemans.fr
[(2)] firstname.lastname@lif.univ-mrs.fr

Abstract

In this article, we present the RATP-DECODA Corpus which is composed by a set of 67 hours of speech from telephone conversations of a Customer Care Service (CCS). This corpus is already available on line at http://sldr.org/sldr000847/fr in its first version. However, many enhancements have been made in order to allow the development of automatic techniques to transcript conversations and to capture their meaning. These enhancements fall into two categories: firstly, we have increased the size of the corpus with manual transcriptions from a new operational day; secondly we have added new linguistic annotations to the whole corpus (either manually or through an automatic processing) in order to perform various linguistic tasks from syntactic and semantic parsing to dialog act tagging and dialog summarization.

Keywords: Corpus, Linguistic Annotations, NLP tasks

1. Introduction

The RATP-DECODA corpus (Bechet et al., 2012) has been acquired in different operational days at the CCS of the Paris transportation center (RATP) for a project funded by the French Research Agency (ANR). The corpus consists of real-life telephone conversations and focuses exclusively on spoken language. These conversations introduce a customer problem to which the agent attempts to provide solutions following a protocol. They have duration variable from 1 to 12 minutes. The number of turns in a conversation and the number of words in a turn are highly variable. The majority of the conversations have more than ten turns. The turns of the customer tend to be longer (> 20 words) than those of the agent and they are more likely to contain out of vocabulary words that are often irrelevant for the resolution of the problem. This corpus gives an illustration of what the actual interactions between customer and user in an information service.

All the conversations of the RATP-DECODA corpus have been manually transcribed. In addition, manual annotators have given a theme to each conversation from a list of 14 labels (13 semantic classes and a *NULL* class). From these manual annotations, several additional layers of linguistic annotations have been added, such as disfluencies, syntactic and semantic parses, discourse structure... These annotations have been done either manually, or through an automated process with or without human supervision.

This paper recalls all the annotation available on the corpus through a presentation of all the linguistic tasks that can be carried on it.

2. The RATP-DECODA corpus

Data mining is an area of research which interest is growing. In addition, the recent increase in call-center and the low cost of storing audio data allowed the recording of oral messages and building of databases very large. These call-centers are a challenge for businesses and can help answer two important applications of the research area, which is the "Speech Analytics": the analysis of large corpus of dialogues to allow to diagnose problems encountered or the extraction of knowledges and the analyse of the behavior of their users. All these fields of application require manual annotations of large volumes of data to enable the establishment and adaptation of ASR System but also recognition and classification models.

The goal of the DECODA project is to enable the development of these strategies and these applications while allowing the annotation effort to be less important.

However, the type of speech that is found in the conversations of call-centers is quite far from what usually treat ASR Systems. In conversations analyzed here, regardless of the single relationship between user and customer, there is no professional speakers. Speakers can be all ages, no-native, can speech with an accent, can cry or scream.... Speech has a greater degree of spontaneity yet. The difficulties are increased by: hesitation about corrections, discourse markers, ungrammaticality, noises. However, a priori knowledge available on the semantic content of messages, particularly through specific documentation of the application and codified scenarios can help overcome some of these difficulties.

2.1. Characteristics of the previously released corpus

The first version of the RATP-DECODA corpus contains 1609 conversations collected in the RATP services on two days. The first day is of November 12, 2009 and the second is of 06 December 2010. Note that this corpus has been acquired in winter, during periods of high traffic in the capital. Similarly, one of the two days corresponds to a day strike in the Parisian metro. Thus, the most common theme concerns the traffic and conversations are mostly noisy since many users call a station not seeing their bus or Metro come.

Each conversation has been manually transcribed and annotated in terms of the dominant theme. 8 major themes were initially selected from a list of 22 proposals. This list have been provided by the RATP call-center itself. For conversations that may contain many themes, only one theme is manually annotated as dominant based on rules contained in the service documentation.
However, this first version is limited by the lack of thematic diversity. The addition of a third collection day was required to extend the types of Telephone conversations and the discussed themes. The third day, the July 4, 2011, allows collecting conversations at the start of summer holidays. The issues are so different: they involve more the new schedules and changes of itineraries related to the repair work.

2.2. 500 new Conversations

This new day was already manually transcribed but we still had corrected the transcripts: the main objective was to verify the alignments on the audio signal. Short sentences to make a title for each conversation and summaries were also made as for the conversations of other two days. Some words like "looking for an itinary," "lost of wallet" or "traffic accident with a bus" used to provide the main theme and a brief summary complete this "semantic" work. In addition to the 8 themes originally used that corresponded to the most commonly viewed topics in the first state collection, it became necessary to resort to the use of less common themes from the list outlining the problems faced by the RATP (1). 4 themes have been added: RETT for claims as a result of strikes or for a mistake in purchase cards, CPAG for altercations with drivers or traffic accidents, OFTP for the possibilities of alternative transport in case of repair works (development of shuttles) and JSTF for requests for proof of delay.
The 4 themes are certainly not the most popular but they guarantee greater coverage. Then they meet the needs of the most specific conversations. Thus, the CPAG theme is relatively rare in the corpus but gives a theme in conversations that indicate a problem with an agent of the RATP (troubles, wrangles...). Before the entry of this theme in the set of theme labels, relevant conversations labeled remained NULL.
Also, we realized that many conversations were labeled NULL, while others hold the label TRASH. It was useless to keep two labels to describe conversations unrelated thematic possibilities and out of context. So we decided to unify this category "out-of-domain" keeping only the NULL label and the propagating all relevant conversations. A new review has helped to unify the annotations.

2.3. The corpus today

The RATP-DECODA is composed by two different parts: in the one hand, manual transcriptions and semantic annotations in themes of Telephone conversational speech with the audio and, in the other hand, automatic transcriptions and multi-level annotations.
The RATP-DECODA corpus is composed by 2109 conversations, which are all a theme annotation received. This represents 67 hours of spontaneous speech without overlap

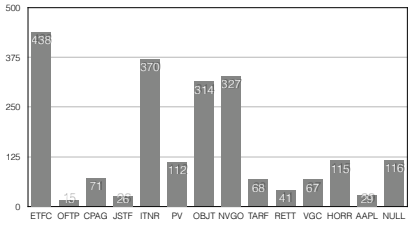

Figure 1: Distribution of the 13 themes of the RATP-DECODA corpus + NULL

for 104 hours of audio. By using the 22 themes from RATP documentation, 13 themes have been selected to cover all the issues arising from conversations. Thus, conversations labeled ERROR in the first two collection time were taken up and were able to receive, if applicable, a theme.
Finally, a last case of theme annotations concerning conversations that deal with the loss of a Navigo card. It appeared that these conversations were not getting exactly the same treatment: some were labeled OBJT to match the theme "loss of an object" while others received the "NVGO" theme when the annotator saw we were talking before all of the Navigo card. Unification was undertaken and all relevant conversations were relabelled in NVGO.
The Decoda corpus is unified today and provides a breakdown into themes that can perceive the divesity of the problems addressed in a call center of a transport company. The 13 selected themes are presented in table 1 adding the theme NULL for conversations "out-of-domain" or are errors.

	Theme
Itinerary	ITNR
Lost and found objects	OBJT
Time schedules	HORR
Transcription cards	NVGO
Cards for specific categories	VGC
Traffic state	ETFC
Fares	TARF
Fine	PV
Special offers	OFTP
Problem with an agent	CPAG
Justification of delay	JSTF
Repayment	RETT
Railway related problems	AAPL
Out of domain conversations	NULL

Table 1: Conversation themes

As can be seen and as might be expected (Figure 1), the themes ETFC and ITNR constitute a majority. Then the OBJT and NVGO themes follow. These 4 themes are central to the problems of an urban transport company. The added four topics constitute a minority but allow not to see the part of conversations which have received the NULL

label become too important.
Finally, it should be noted that some conversations are problematic insofar as they deal with several themes. The user, having asked a first question and heard the answer, allows himself to speak about a second problem. To overcome this difficulty, we used a special label in our annotations inspired by the Theorie of Dialogue Acts (Mann and Thompson, 1988) and (Bunt, 1995). This label is "NEXDEMANDEUSAGER" and placed in a conversation about the first segment of a customer, when he makes a new demand.

3. Linguistic annotations for performing NLP tasks

In addition to the main theme, every conversation has been the subject of several levels of linguistic annotation. In the first version of the corpus, syntactic annotation was added through a semi-supervised process (Bazillon et al., 2012). Since then, several other layers have been added such as semantic frame annotations (Trione et al., 2015) or dialog summaries through the SENSEI European Project[1] (Favre et al., 2015). In addition to previous annotations, we present in this section all the new annotations that have been done, either manually (discourse parsing) or automatically (ASR transcription) on the corpus. This presentation is done according to the kind of Natural Language Processing task that can be performed on the corpus thanks to these annotations.

3.1. A corpus for the ASR of very spontaneous conversation speech

All the conversations have been automatically transcribed with the LIUM automatic speech recognition (ASR) system based on the Kaldi decoder (Povey et al., 2011) with DNN acoustic models. This ASR system is derived from the one described in (Rousseau et al., 2014) that generates, for each conversation, a lattice of word hypotheses linguistically rescored by LIUM tools derived from the CMU Sphinx project (Deléglise et al., 2005) to provide the most likely word sequence hypotheses.
The ASR language model has been adapted to the application domain. To get both automatic transcriptions as accurate as possible and homogeneous performances on the entire DECODA data, a leave-one-out approach was used in order to compensate the lack of data necessary to estimate language models: data were split into three datasets with the same size. Each dataset was transcribed with the ASR system using a language model estimated on the two other datasets and interpolated with a general language model similar to the one used in (Rousseau et al., 2014), built from manual transcriptions of broadcast news speech, articles crawled from Google News, the French Gigaword corpus, articles from the "Le Monde" newspaper and from Wikipedia. The interpolation weights between the in-domain language model (0.85) and the general one were the same for each dataset. The vocabulary of the ASR system contains 159,515 words, including all the words present in the DECODA corpus.
The global ASR word error rate (WER) is 34.5%. This high error rate is mainly due to speech disfluencies and to adverse acoustic environments for some conversations when, for example, users are calling from train stations or noisy streets with mobile phones. Furthermore, the signal of some sentences is saturated or of low intensity due to the distance between speakers and phones. The leave-one-out approach used to transcribe the entire corpus minimizes the bias related to the use of these automatic transcripts for NLP experiments.

3.2. A corpus for the syntactic analysis of spontaneous speech

As described in (Bazillon et al., 2012), the first version of the corpus has been annotated with Part-Of-Speech, disfluencies, named Entities and syntactic dependencies relations. This is a partially automatic process. Since the first release these annotations have been normalized through the ORFEO projecti[2] (Nasr et al., 2014). At the current state, this corpus represent a unique opportunity to study the syntactic phenomenons occurring in call-center dialogues in French.

3.3. A corpus for the semantic frame parsing of spontaneous speech

The *RATP-DECODA* has been annotated with semantic frames as presented in (Trione et al., 2015). We used a FrameNet-based approach to semantics that, without needing a full semantic parse of an utterance, 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 ?*). 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. In the RATP-DECODA corpus, 188,231 frame hypotheses from 94 different frame definitions were found. We decided in this study to restrict our model to the frames generated by a verbal lexical unit. 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). 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.
In the released corpus, the semantic frame annotations are projected at the word level: each word is either labeled as `null` if it is not part of a frame realization, or as the name of the frame (or frame elements) it represents. In our corpus, 26% of the words have a non-null semantic label and there are 210 different frame labels. A lot of ambiguities

[1] `http://www.sensei-conversation.eu`

[2] `http://www.projet-orfeo.fr`

[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

come from the disfluencies which are occurring in this very spontaneous speech corpus.

3.4. A corpus for the discourse parsing of conversational speech

For discourse parsing, we developed a model inspired by (Mann and Thompson, 1988). RST considers a coherent text made of segments containing elementary discourse units (EDU) expressed by clauses and related by a small set of discourse relation types. This set of discourse relation types provides a structured view of the conversation and of the turn between user and customer. A typical conversation analysis task of customer care service (CCS) applications is to process conversations between a customer and an agent who attempts to follow a protocol defined in the application domain description. Based on the application requirements, a report has to be compiled for each conversation.
Apart from the theme annotations that allow categorizing conversations, verifications that have unified the entire corpus, annotations in Dialogs Acts from a free adaptation of the theory of Bunt (Bunt, 1995) have given structures for each conversation. The aim was twofold: on one hand, digressions and theme changes must be detected. On the other hand, the objective was to ensure that the most relevant elements to complete the report and determine the theme were located at strategic moments of dialogue.

3.5. A corpus for evaluating abstractive methods for dialogue summarization

Supervisors in the call-center used to write a small summary of all the conversations they have monitored. These summaries, that we call *synopsis* contain a very abstractive view of a conversation, summarizing in just a few sentences at the same time the main topic, the interaction and the final result. These summaries are a great opportunity to evaluate abstractive summarization methods of human-human dialogues. A pilot task was proposed at Multiling 2015 (Favre et al., 2015) in this direction, on the French RATP-DECODA corpus and on a small portion of it translated to English.

4. Conclusion

In this paper, we presented the improvements we made to the first version of the RATP-DECODA corpus. We first reported the addition of new conversations in this corpus of telephone conversations. Then we presented the new linguistic annotations added, as well as a list of NLP tasks that can be performed thanks to this corpus. It will be soon available in its totality on line at `http://sldr.org/voir_depot.php?id=847&lang=fr&sip=1`.

5. Bibliographical References

Thierry Bazillon, Melanie Deplano, Frederic Bechet, Alexis Nasr, and Benoit Favre. 2012. Syntactic annotation of spontaneous speech: application to call-center conversation data. In *Proceedings of LREC*, Istambul.

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Bèze, Renato De Mori, and E. Arbillot. 2012. Decoda: a call-center human-human spoken conversation corpus. In *International Conference on Language Resources and Evaluation (LREC)*.

Harry Bunt. 1995. Dynamic interpretation and dialogue theory. *The Structure of Multimodal Dialogue*, 2:139–166.

Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin. 2005. The lium speech transcription system: a cmu sphinx iii-based system for french broadcast news. In *Interspeech*, pages 1653–1656.

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *Sigdial*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *The Prague Bulletin of Mathematical Linguistics*, 8:243–281.

Alexis Nasr, Frederic Bechet, Benoit Favre, Thierry Bazillon, Jose Deulofeu, and Andre Valli. 2014. Automatically enriching spoken corpora with syntactic information for linguistic studies. In *International Conference on Language Resources and Evaluation (LREC)*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December.

Anthony Rousseau, Gilles Boulianne, Paul deléglise, Yannick Estève, Vishwa Gupta, and Sylvain Meignier. 2014. Lium and crim asr system combination for the repere evaluation campaign. In *Proceeding of the 17th International Conference on texte, Speech and Dialogue*, September.

Jeremy Trione, Frederic Bechet, Benoit Favre, and Alexis Nasr. 2015. Rapid framenet annotation of spoken conversation transcripts. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*.