Proceedings of the LREC2020 Industry Track , pages 6–9
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Promises and Disappointments of Semantic Analysis of Speech-To-Text Applied to Call Center Conversations in an Industrial Setting

Ruslan Kalitvianski, Emmanuelle Dusserre, Muntsa Padró
ELOQUANT
5 allée de Palestine, 38610 Gières
{ruslan.kalitvianski, emmanuelle.dusserre, muntsa.padro}@eloquant.com

Abstract

Recent progresses in speech-to-text technologies, and the marketing hype they create, lead to believe that speech-to-text has become robust enough for reliable semantic analysis of textual transcripts of phone conversations to have a chance to be feasible. This paper describes our experience as a mature provider of semantic analysis of written text in French with analysis of uncorrected machine transcripts of real-life call center two-speaker conversations. We argue that, although much progress has indeed been made, many difficulties remain before French conversation transcripts can be exploited to their full potential.

Keywords: speech-to-text, semantic analysis, opinion mining, classification, call center

1. Introduction

Our team has been developing commercial solutions for semantic analysis for text in French for many years. We propose machine learning and expert-based systems for generic and bespoke multilabel classification, opinion extraction and topic modelling. Since the acquisition of our start-up by a company specialized on customer relations management solutions, the services we provide are shaped on the direction of extracting information from the different channels we propose to our client companies to communicate with their customers.

One of these channels, and probably the most important today for us, is the voice channel. Our company offers a platform for managing call centers with multichannel interactions (phone, emails, chat…), phone calls being nowadays more than 80% of the communications managed. Thus, the amount of data generated from the conversations of our customers with their costumers is huge. To enhance the service we give to our customers, it is a natural step to experiment with the use machine analysis of conversation recordings to automatically determine which topics are being discussed, how the conversation is evolving, what are the emotions expressed by the customer, etc.

There is a profusion of providers of speech-to-text (STT) solutions on the market for the French language. Nuance, Vocapia, Bertin IT, Allo-Media, IBM Watson, Google Speech-to-Text, to name some. A few are specialized in phone conversations, but most offer various models trained for different purposes.

We are aware of numerous works on deep learning-based architectures that have considerably advanced the state of the art in speech transcription, but in this paper, we are interested in mature commercial solutions, being agnostic of the underlying technologies.

This paper describes our experience as an experienced provider of semantic analysis for French with analysis of uncorrected machine transcripts of real-life call center two-speaker conversations. First, we describe the rationale that a company such as ours could have for using semantic analysis of machine transcripts, then we describe the difficulties and obstacles we encountered when trying to apply our technology on this data.

2. Promises of STT: Massive analysis of an abundant resource

Call centers abound with recordings of conversations, which represent a big wealth that companies could exploit to better serve their customers and improve the efficacity of their contact centers. Nevertheless, these data are not as readily exploitable as written text, therefore good STT systems are crucial for the development of these applications.

In what follows we present some interesting tasks that could be performed with a quality speech transcription.

2.1 What STT providers can do

Other than raw speech-to-flow-of-words transcription, many STT systems provide other features:

- "Speaker diarization is the task of determining "who spoke when"" (Anguera et al., 2010). Most commercial STT providers include automatic or semi-automatic identification of the number of speakers in a conversation and attribution of speech turns, performing a kind of speech segmentation, which is useful when the audio is single channel (mono). This task may also involve speaker gender identification.
- Word timestamping and word transcription confidence scores: some systems (Vocapia, Bertin IT) assign a timestamp to each transcribed token. This allows to align tokens with the audio, a useful feature when developing an interactive audio player that displays the transcription. More expert users may benefit from the confidence score ∈ [0, 1] that these systems assign to each token. Other STT providers (Nuance) can output a lattice of transcription candidates which allows NLP experts to select the best transcription path using a custom language model.
- Several systems (Nuance, Vocapia, Bertin, Google STT) can incorporate custom vocabularies for correct transcriptions of persons' names, product names, places, etc. Some accept a simple list of words, other allow to specify a category for each custom word (product name, locality, etc).
- A few systems (Google, Vocapia) have begun adding automatic punctuation to their output, which renders the transcript more natural to read.

2.2 Indexation

The most basic use for speech transcripts is term indexation of the audio. This allows searching among and within the conversations for mentions of specific terms. This is a need that has been expressed by some of our customers during our brainstorming workshops.

2.3 Semantic analysis

We provide Web platforms that allow to graphically display results of semantic analyses on an interactive dashboard (Dusserre *et al.*, 2020). One can thus combine results of different analyses to select, for example, opinions expressed about one or several products or topics. The purpose of these platforms is to provide our clients with an overview of the vocal interactions between their agents and their customers, since the volumes of the data, coupled with the time it takes to listen to a call, vastly exceed the humanly analyzable. Thus, a more ambitious use is machine analysis of the transcripts for:

- Speaker role identification: based on the diarization performed by STT, determine who of the speakers is the client and who is the agent. In a general setting, one cannot systematically say, for instance, that the first person to speak is the agent that is replying to an incoming call. An analysis of what is said is necessary to determine who is who in the conversation. This will allow both filtering in the Web dashboards, and analyses tailored to each type of speaker.
- Call topic extraction: extract the topics addressed in each call, either for statistical analysis over large collections of calls (to determine the most frequent topics and their diachronic evolution) or to classify the call (i.e. for indexing, routing, etc.). This task can be approached in two ways:
  i. Unsupervised topic modelling: given a significant amount of calls, extract the most frequent topics that are relevant to the current domain. This allows for an evolutive modelling of the calls for each costumer and to discover new topics.
  ii. Supervised classification: given a predefined set of categories, use a classifier to determine the topics of a call. This allows to track over time a set of categories that we know are important for the customer/domain.
- Opinion mining: extract feedback given by the customer (positive or negative opinions, action requests, threats, etc.). This would allow to perform real-time alerting about unhappy customers, study the opinions related to each topic or category, etc.
- Domain-specific named entity recognition: detect the mentions of a given product or specific entity to gain insights into its popularity, understand the satisfaction level related to it, etc.
- Automatic summary generation and action item detection: this has been attempted by the CALO Meeting Assistant Project (Tur *et al.*, 2010) for English, and, more recently by the REUs project for French (Patel *et al.*, 2019).

2.4 Good/bad practice identification

Our client companies' agents say that the mere fact of reading one's own conversation transcript al-lows them to better themselves by an a posteriori analysis of how an interaction.

An advanced semantic analysis may also help identify the specific rhetoric used by the agent that allowed him or her to close a deal, or, on the opposite side, to see what went wrong in an interaction.

Many companies have scripts that their agents must adhere to (for instance mentioning a product's name N times during an advertisement), and semantic analysis could be used to check for script conformity.

2.5 Combining NLP extracted information with structured data

Information extracted from audio can be combined with that extracted from other data sources, which can be especially useful in a multichannel customer relation management platform. Some examples of other data can be event information (how long was the conversation, did the customer already contact the company before, etc.), customer information (age, gender, etc.), closed questions answers (e.g. satisfaction note the customer gave in a survey), among others. Thus, we can imagine several applications of combining these kind of data with data extracted from conversations: profiling of customers and agents to assign them to the best suited agent, prediction of satisfaction rate given the contents of the conversations, computing automatically first call resolution rate, etc.

3. Difficulties

Our first attempt at semantic analysis of conversation transcripts was for a client company that routinely recorded their pre-sale and customer support conversations via the several call center platforms they employ. These are conversations recorded in France, whose length ranged from less than a minute to over 40 minutes long.

We agreed to transcribe conversations that were between 1 and 10 minutes long and transcribed over 10000 such audios over a period of one year. All audios are single channel, 8KHz, 64 Kbit/second, encoded with G711.

3.1 Choice of STT provider

The first obstacle was selecting the best provider of STT for French. We had to construct a gold standard of transcription on the data of the client, a process that is very time-consuming: on average, transcribing 4,5 minutes of audio required an hour of work per person.

The task was distributed to 20 participants which allowed us to collect 90 minutes of transcribed speech. Instead of making our participants work from scratch, we chose to give them machine transcripts from a promising STT tool, that they would correct via a text editing tool. *A posteriori*, all participants felt that having a machine "pre-transcript" was useful, and accelerated the transcription process.

3.1.1 The issue of evaluation

We chose to evaluate several systems using the word accuracy metric (WAcc = 1 - Word Error Rate[1]). We normalized input texts by stripping punctuations (if any), lowercasing words, removing some disfluency markers

[1] https://www.vocapia.com/glossary.html

("euh", "hum", etc.) as well as newlines. The texts to be compared are therefore streams of words. We chose not to evaluate speaker diarization, because was of lesser priority.

We evaluated a total of ten models of five leading STT providers. Given that the conversations centered on the client's products (heating appliances), we built a lexicon of client-specific terms, and used it as a parameter for systems that accepted vocabulary customization.

On twenty audios of 4.5 minutes each, the lowest observed average word accuracy rate was 0.47 and the highest was 0.73. The median WAcc was 0.696. The highest average WAcc was obtained with a system that used a phone conversation model supplemented with the client's vocabulary.

It is important to note that merely focusing on average performance occults the performance of the system in the best and worst cases, and, more generally the variability in performance. Thus, the system with the best average WAcc transcribed only 10% of the audios with a WAcc at or above 0.8, whereas the second-best transcribed 15% of the audios with a WAcc at or above that threshold, but conversely it produced less transcriptions that reached the 0.7 WAcc threshold.

3.1.2 Lexical fidelity vs task usefulness

Lexical fidelity may not be the best or the only measure of the performance of a STT system. We decided to perform and extrinsic evaluation by comparing the results of our analyses on the gold standard transcriptions with those on the automatic transcriptions using the two best STT providers with their best configuration. We evaluated three analyses that produce sets of words or categories:

- Concept extraction, which extracts nouns and nominal groups that are significant to the domain using a terminology extraction system based on Sclano and Velardi (2007).
- Categories, which are client-specific themes, based on a ML classifier (which uses as features the word, the lemma, noun phrases, and semantic annotations coming from domain-specific gazetteers), and hand-written rules based on the TokensRegex formalism (Chang and Manning, 2014).
- Domain-specific Named Entities, which are simply a list of our client's products and their synonyms.

Table 1 demonstrates that a system that performs better at transcription (average WAcc = 0.73) may reveal itself to be less useful for semantic analysis than a system with a slightly worse performance (average WAcc = 0.699).

		F-measure		
		Concepts	Categories	Domain NEs
Best system	Average	**0.51**	0.79	0.69
	Median	**0.52**	0.85	0.73
Second-best system	Average	0.50	**0.86**	**0.83**
	Median	0.48	**0.92**	**0.86**

Table 1: performance of three semantic analyses on machine transcripts by the two best systems (as determined by their average WAcc), with human transcripts analyses as reference.

These results show that, paradoxically, analyses performed on the output of the best system show less agreement with analyses of human transcripts of the same audios than analyses of the output of the second-best system. The discrepancy is likely due to differences in the abilities of these systems in integrating customer-specific vocabulary into the transcription model. More generally, that shows the importance of performing this kind of evaluation on top of core task evaluations, since they may reveal that the most useful systems for industry purposes may not be the ones that have best scores on paper.

3.2 Insufficient audio quality

Real-time telephony is a low-latency transmission whose only requirement is human intelligibility of speech. Signal is deformed at many steps, including capture (low quality microphones), encoding, transmission (loss of packets), mixing (fusion of two speaker channels into one).

Moreover, the system we selected based on best average performance does not attempt automatic punctuation, which results in a flow of words that seem odd when read, even if correctly transcribed, partly because they lack prosodic features such as pauses.

Thus, the indexing and reading promise of STT appears only partially fulfilled. Reading a transcript is often a tedious task, because of the reasons above, and because one must guess what words were actually uttered when one reads an erroneous portion, which is a cognitively demanding task.

3.3 Insufficient transcription quality

Upon subjective evaluation of transcription quality, it turned out that a WAcc of 0.7 corresponded to a rather low-quality transcription. Most of the correct words were trivial words, there were portions that were not transcribed, often at the beginning of a speech act.

Moreover, the system we selected based on best average performance does not attempt automatic punctuation, which results in a flow of words that seem odd when read, even if correctly transcribed, partly because they lack prosodic features such as pauses.

Thus, the indexing and reading promise of STT appears only partially fulfilled. Reading a transcript is often a tedious task, because of the reasons above, and because one must guess what words were actually uttered when one reads an erroneous portion.

3.3.1 Speaker diarization

This is a necessary step when one has mono audio (as we did) and wants to build a speaker classification system to determine which speaker is the client and which one is the agent. The two typical errors we observe with speaker diarization are:

- Incorrect identification of the number of speakers: in our context two is the correct number, but sometimes more are found, or, when the voices of the two speakers are similar, only one. Some systems can take the maximum number of speakers as a parameter, others cannot.
- Incorrect speech turn boundary placement, which attributes words to the wrong speaker, making the

transcript less comprehensible and confusing the speaker role identification system.

3.3.2 Impact of noise and errors

As mentioned before, there are many errors on the automatic transcriptions. This means that the semantic analysis systems that we build to extract information have to be very robust to these errors. This implies that it is very challenging to develop rule-based or Machine Learning systems that correctly model a very variable set of contexts. Here, we must deal with the already rich and variable spoken language with noise added by the STT system.

Furthermore, French is especially difficult to analyze because of its high degree of homophony. It is crucial to have a good language model, and to adapt it to each domain, since each customer will have a set of specific words (names of products, for example) that are unknown by the system or that appear in unexpected contexts. In our experience, this lexical and contextual tuning are mandatory to improve the quality and, most importantly, task usefulness of the transcriptions. Adding words is an option for many STT system, but it would also be useful if words that are observed in machine transcriptions and are unlikely in the client's context could be manually removed from the model's vocabulary.

3.4 Analyzing text vs analyzing speech

A difficulty that is not necessarily related to the quality of the transcription is the vast difference between written one-time comments and spoken conversations.

3.4.1 Disfluencies of speech

We use linguist-defined rules that describe specific lexical and syntactic patterns to detect opinions. Spoken language is much less precise than written comments, and it is much harder to write patterns that match discourse filled with disfluencies, interruptions and transcription errors.

3.4.2 Loss of prosody

Much emotional information is conveyed by prosody, which could be used to increase confidence in opinions detected in the text. Thus, to correctly detect opinions, studying not only lexical items, but also the prosody seems mandatory. None of the systems we tried supplied this information.

3.4.3 Expressions of opinions

Usual sources for opinion mining are surveys or reviews that are rich on opinionated expressions since this is their purpose. Detecting opinions in an oral conversation is much more difficult, for several reasons:

- As mentioned, a much emotional information is conveyed by prosody, not on the language itself.
- The purpose of many conversations is often to factually discuss topics, not necessarily to express satisfaction or unsatisfaction. Thus, the opinions are found much less frequently, mixed with a lot of other information, and often expressed implicitly.
- Given the oral nature of the speech and the errors from STT systems, the sentences are rarely syntactically complete, thus, opinion extraction systems based on syntactic patterns only rarely match.
- Expressions such as "good", "very well" or "ok" often imply an opinion in written reviews but are used as mere acquiescence in phone conversations.

4. Conclusion

As a call center management software experts and NLP experts, we are highly interested on exploit-ing STT solutions to process the big amount of data that our customers produce. We believe that the call center software of the future will allow companies to better serve their customers by automatizing as much as possible the study of the needs, expectations, and satisfaction of the customers.

The promise of STT systems is to accurately transform audios into written texts, which would al-low their browsing, indexing and detailed analysis.

In this paper, we have presented some of the applications that we can imagine of such a transcription and analysis, but also the current limitations discovered in practice when applying current state-of-the-art commercial solutions to our real-world data. Namely, even if we can extract some information as the topics of a conversation, the quality of the transcription is often not good enough to be read without listening to the audio.

In our domain, we observed two leverages that can improve the quality of STT systems: the quality of the audio, which in our case is limited by the phone lines, and the use of in-domain lexica to tune the STT models. These are, of course, known issues for researchers, but what is sometimes ignored is the difficulty in real life to obtain (or create) audio in good quality and good lexical resources.

The lessons learnt from our experiences is that big attention has to be given to these factors before setting up a project, and that there is still a long way to go to be able to fully analyze call conversations, but we remain optimistic about the possibility of reaching this goal and to industrializing se-mantic analysis of this channel.

5. Bibliographical References

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356-370.

Chang, A. X., & Manning, C. D.: TokensRegex: Defining cascaded regular expres-sions over tokens. *Technical Report CSTR 2014-02*. Department of Computer Science, Stanford University (2014)

Dusserre, E., Kalitvianski, R., Ruhlmann, M., & Padró, M. (2020). Analyse sémantique de transcriptions automatiques d'appels téléphoniques en français. *Actes de la 6e conférence conjointe JEP, TALN, RÉCITAL, Volume 4: Démonstrations et résumés d'articles internationaux.*

Patel, N., Lannes, M., & Pradel, C. (2019, July). Patrons linguistiques pour l'extraction de tâches dans des transcriptions de réunions. In *Actes de PFIA 2019, pp. 158–166*. PFIA.

Sclano, F., & Velardi, P.: TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In: R. J. Gonçalves, J. P. Müller, K. Mertins, M. Zelm (Éd.): Enterprise Interoperability II, pp. 287-290. Springer London (2007)

Tur, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tur, D., Dowding, J., ... & Frederickson, C. (2010). The CALO meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1601-1611.