



On the Use of Linguistic Features in an Automatic System for Speech Analytics of Telephone Conversations

Benjamin Maza¹, Marc El-Beze¹, Georges Linares¹, and Renato De Mori^{1,2}

¹LIA, University of Avignon, France

²McGill University, School of Computer Science, Montreal, Quebec, Canada

{benjamin.maza, marc.elbeze, georges.linares, renato.demori}@univ-avignon.fr

Abstract

A research on the analysis of human/human conversations in a call centre is described. The purpose of the research is to provide short reports of each conversation with information useful for monitoring the call centre efficiency. Data from real users discussing over the telephone with agents are processed by an automatic speech recognition (ASR) system. Reports are grouped into classes by the agents based on predefined taxonomy. A train set of manually transcribed data is used for training the extraction of features relevant to the application and the classification of the conversations. The use of all the words of the application vocabulary, of automatically selected keywords, and of automatically learned sentence chunks containing semantic classes of words are compared and evaluated with a totally different test set. The results show a significant increase in performance when chunks are used even in comparison with the use of bags of words obtained with a boosting algorithm.

Index Terms: speech analytics, human/human conversation analysis, call centre performance monitoring, speech understanding, and dialogue classification

1. Introduction

A number of activities related to the analysis of human/human conversations are referred to in the literature as *speech analytics*. They are reviewed and discussed in [6]. The purpose of these activities is to extract features of the conversations useful for various purposes. Among them, the activity considered in this paper is the compilation of reports about call centre telephone dialogues between agents and users for solving specific user problems.

Useful information for inferring the required features can be words, syntactic structures expressing semantic knowledge, dialogue and discourse structures relevant to the application purpose. Other approaches perform information retrieval (IR) of spoken messages with operations based on a vector space model (VSM) [3].

Attempting to do speech analytics with Automatic Speech Recognition (ASR) is made difficult by the fact that ASR systems are not perfect. ASR errors are frequent when automatic transcriptions are obtained from speech of many casual, unknown users speaking on the telephone in a large variety of acoustic and noisy environments. Furthermore, in applications of this type, a large variability can be observed in the expression of the same concept. Errors due to the above mentioned degrees of variability can be tolerated in applications in which relevant concepts are frequently expressed many times in the same dialogue. This is the case considered in the system described in this paper that is conceived to fulfil important call centre management needs. The application scenario is that of an agent

that follows a well defined dialogue protocol in interacting with a casual user. The purpose of the dialogue is to solve a specific problem, such as informing about an itinerary or discussing a complaint. The data have been collected in a call centre of the RATP public transportation system in Paris. The task is that of producing conversation reports for which a set of concepts can be defined. These concepts are expressed in dialogue turns in which other concepts, not relevant for the application, are also expressed. For example, if the problem is the request about a transportation mean for a given itinerary, what is essential for accomplishing the analysis task is to detect that the problem is described by two related concepts, namely a request and an itinerary. The details of the itinerary and other factual information are irrelevant. Furthermore, dialogues are between an agent that attempts to follow a pre-defined dialogue structure and a customer whose behaviour is unpredictable. The agent solicits information about the details of the problem, receives answers, may ask for confirmation and call other services for advice. The user may require explanations of the solution or the repetitions of parts of it. Dialogues of this type contain frequent redundancies making it possible to automatically understand the type of dialogue even in presence of ASR errors.

The research described in this paper compares and integrates solutions based on pure IR methods with solutions based on the automatic characterization and use of semantic, application relevant, information. Automatic learning and detection of sentence chunks characterizing semantic information is considered and evaluated. A specific confidence indicator is introduced, making it possible to increase the percentage of successfully completed tasks by selecting a small proportion of conversations to be manually processed.

Section 2 discusses related work. Section 3 introduces the system architecture and the types of features used for producing conversation reports. Section 4 describes experimental results showing a significant increase in performance when chunks expressing semantic features are used instead of pure lexical features.

2. Related work

IR approaches to spoken language understanding (SLU) with considerations relevant to conversation analysis are reviewed in [3]. Other approaches use features that can be extracted performing full parsing, shallow or partial parsing or with a process that does not require any type of parsing. Parsing techniques for extracting dialogue features are reviewed in [10]. Approaches that are somehow related to the one proposed in this paper are now briefly discussed.

In applications involving casual speakers that may not always follow grammar rules for forming their utterances and likely to have a non negligible amount of ASR errors, it is not advisable to use parsing features. Instead, it is interesting to

consider using chunks of words that are relevant for the concepts to be hypothesized in the application.

In [7], a general-purpose approach based on word dependency is proposed in which dependencies are hypothesized independently of tasks/domains. A maximum entropy method with multiple features, such as word class, word position, is employed for training a model. The dependency structure of an unseen word sequence is determined so that pairs of words in dependent relationships give the maximum total probability.

Recent review on keyword and chunks unsupervised construction can be found in [2]. These features are used for indexing and retrieving spoken lectures. Three different sets of features are proposed, namely: prosodic features, lexical features and semantic features. Prosodic features are *Duration Related*, *Pitch Related*, *Energy Related*. Lexical features are TF-IDF. Term frequency (TF), inverse document frequency (IDF), and TF-IDF are computed for each term, for variations of left and right context of a word, and for parts of speech (POS). Semantic features are obtained with probabilistic latent semantic analysis (pLSA) assuming that each document contains a set of hidden topics and the probability of a term given a document can be expressed as a summation over topics of the probability of the term in a topic times the probability of the topic given the document. Latent Topic Significance (LTS) and Latent Topic Entropy (LTE) are defined for each term. Key phrases are extracted using branching entropy computed on a data structure for strings of symbols called Patricia tree (PAT tree).

Extraction of chunks in French from written text has been recently described in [4].

New types of automatically formed chunks suitable for the type and purpose of the application considered in this paper are introduced in the following section. They are obtained with a discriminative criterion that attempts to make evident only sequences of words and semantic word classes that are relevant for separating application relevant concepts from other concepts expressed in a conversation and for composing a report for each conversation containing all and only the information useful for the analysis task.

3. System architecture and features

The features introduced in this section are used in the system architecture scheme shown in Figure 1. Spoken conversations are processed by an ASR system at the output of which features are extracted for inferring components of a conversation report. The report is stored in an archive. For each report, parameters used for monitoring the performance of a call centre service are computed. If the report cannot be obtained automatically with sufficient confidence; then the report is compiled by a human expert. In order for the service to be useful, automatic report generation has to be reliable and the rate of manually generated reports should be low.

As conversation reports to be compiled belong to a finite number of types, feature selection can be performed in a classification framework. Ten report types, referred to in the following as *report classes* have been identified as containing enough information for obtaining measures useful for monitoring the performance of the call centre where the data were collected. These classes belong to the domain of urban transportation. As the ASR word error rate (WER) can be higher than 50% for real-life conversations between agents and casual users, parsing the ASR results is unreliable, thus parsing features were not considered.

Three types of features were evaluated, namely, all the words of the ASR vocabulary, keywords and chunks extracted with automatic discriminative methods having the objective of maximizing the correct classification of a train set. Features were compared using the same classification method using a cosine similarity measure between a vector of feature parameters for a report class and a similar vector of parameters computed for each conversation of a test set. Part of the train set was used as development set for tuning classification as the WER is 45% on the train set because the acoustic models of the ASR system were adapted with a limited amount of available data starting with models trained with telephone data of other corpora.

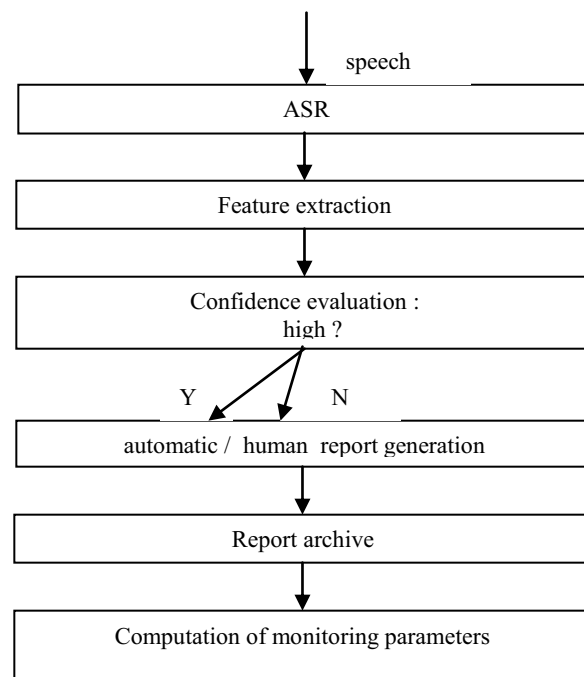


Figure 1 – system architecture

The ASR system being error prone, it is important to consider features that, even if partially recognized can be useful for performing a correct classification by exploiting redundancies in the conversations.

Let $RC = \{r_1, \dots, r_m, \dots, r_M\}$ be the set of report classes, $V = \{w_1, \dots, w_n, \dots, w_N\}$ be the whole ASR vocabulary and $V_{kw} = \{kw_1, \dots, kw_k, \dots, kw_K\}$ be a vocabulary of keywords that are the seeds used for inferring sentence chunks. Some words in vocabulary V are grouped into semantic classes such that *bus_stop*, *metro_station*, *lost_object* obtained from domain knowledge. These word classes are considered as words in the keyword selection and chunk formation described in [4] and summarized below.

$\forall r_m, w_n$ compute the class probability $P(r_m | w_n)$,

$\forall w_n$ compute the *coverage* $\gamma(w_n)$ of the train set defined as follows:

$$\gamma(w_n) = \frac{\text{number of conversations containing } w_n}{\text{total number of conversations}}$$

The purity of w_n is defined by the Gini measure :

$$G(w_n) = \sum_{m=1}^M P^2(r_m | w_n)$$

Word w_n becomes a keyword

$w_k \in V_{kw}$ if $G(w_n)$ and $\gamma(w_n)$ are both above thresholds

determined with the development set with the objective of maximizing correct classification of report classes.

Let $C = \{c_1, \dots, c_j, \dots, c_J\}$ be a vocabulary of chunks obtained with an iterative procedure that starts from a keyword $w_k \in V_{kw}$ and sticks it together with an adjacent word $w \in V$ to form a *term* t_ℓ if the growth of the purity $G(t_\ell)$ and the coverage $\gamma(t_\ell)$ are above thresholds determined by maximizing classification accuracy of the development set. Notice that disfluencies are ignored and do not prevent compounding terms in this process that continues in growing terms until purity becomes 1 or coverage goes below threshold. Partial chunk detection is possible resulting in a lower contribution in the computation of the chunk score used for classification as described later on. Appropriate heuristics are used for hypothesizing chunks whose detection partially overlaps in time. Table I, shows the example of a sentence transcription in French, its annotation in terms of chunks in English, the ASR transcription and the resulting automatic hypothesization of chunks. It appears that in spite of ASR errors, chunks (in bold) are correctly hypothesized.

Table I- Example of manual and ASR transcriptions with and without automatic annotation with chunks (in bold)

Manual transcription	<i>je vous appelle pour un renseignement j'ai perdu mon portefeuille euh sur la ligne 4.</i>
Manual transcription with chunks	<i>I call to inquiry because I lost lost-an-object on the bus line-NumLine.</i>
ASR output	<i>des choses tout renseignement oui chercher perdu mon portefeuille <unk> la ligne quatre.</i>
ASR output with chunks :	<i>tout renseignement chercher lost-an-object <unk> monsieur line-NumLine.</i>

Classification is performed by comparing a vector \bar{X}_d of scores for the features of the d-th conversation and the same vector \bar{X}_m computed with the features of all the conversations of a report class $r_m \in RC$ in the train set.

The d-th conversation is classified into class \hat{r} according to the following decision rule:

$$\hat{r} = \arg \max_m \cos(\bar{X}_d, \bar{X}_m)$$

The elements of vector \bar{X}_d are scores whose computation depends on the type of features used. It contains words and is progressively augmented with classes and chunks. For every chunk c_j , there is an element $x_{d,j}$ in \bar{X}_d defined as follows:

$$x_{d,j} = \begin{cases} \sigma_j & \text{if } c_j \text{ fully hypothesized} \\ i_{j,f} & \text{if only } f \text{ words of } c_j \text{ are hypothesized} \end{cases}$$

where σ_j is the score computed on the fully detected chunk c_j and $i_{j,f}$ is the score computed on the f-th corrupted version of chunk c_j and is proportional to the number of correctly detected components of the chunk. The logarithm of score σ_j is a log-linear combination of the term frequency (TF), where terms are chunks in this case, and the Gini measure $G(c_j)$. The

same type of computation is performed for partial chunks detection using appropriate scores. The inverse document frequency (IDF) used in other approaches has not been used here because the corresponding exponent estimated with the development set was found to be close to zero.

4. Experiments and results

A corpus of 665 telephone conversations between agents and users was collected at the call centre of the RATP public transportation service in Paris. The corpus was split into a train set of 512 conversations and a test set of 153 conversations with a total of 445114 words. Part of the training corpus (128 dialogues), has been used as development set to tune the system thresholds. Conversations belong to ten dialogue types involving the speech acts *inquiry* about concepts like *itinerary*, *lost and found*, *time schedules*, *strikes*, *state of the traffic*, *delays*, *transportation tickets*; *cards* and two types of *complains*. Their frequency in the test set varies between 3% and 26%. Each conversation is manually transcribed and annotated with the report provided by an agent. These reports can be classified according to the 10 dialogue types. Reports were manually analyzed and corrected or completed if necessary.

The ASR system used for the experiment is the LIA-spectral system [5] with 230000 gaussians in the triphone acoustic models. Model parameters were estimated with maximum a-posteriori probability (MAP) adaptation of 150 hours of speech in telephone bandwidth with the data of the train set. The vocabulary contains 5782 words. A 3-gram language model (LM) was obtained by adapting with the transcriptions of the train set a basic LM. An initial set of experiments were performed with this system resulting with an overall WER on the test set of 58% (53% for agents and 63% for users). These high error rates are mainly due to speech disfluencies and to adverse acoustic environments for some dialogues with users calling from train station 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.

Using the classification method described in the previous section, the results reported in Table II were obtained. The term *reference* is used to indicate manual transcriptions, while the term ASR indicates the use of automatic transcriptions obtained by the ASR system. Classification accuracy is defined in this case as the ratio of correctly classified conversations into one of the 10 types of the annotated reports. It appears from Table II that the use of vectors \bar{X}_d integrating chunks with words systematically leads to higher performance compared to the use of words and semantic word classes. Particularly significant are the improvements with respect to a pure IR approach that uses all the vocabulary words as features. Notice that the classification method is the same for all feature types. For the sake of comparison, an approach based on the AdaBoost learning algorithm [8] was used to infer bags of words for each conversation class. With this approach a classification accuracy of 110/153 was obtained when training and testing on the reference transcriptions as opposed to 133/153 obtained with the approach proposed in this paper. Furthermore, an accuracy of 98/153 was observed with the Adaboost training approach using manual and ASR transcriptions while an accuracy of 125/153 was obtained with the approach proposed in this paper.

Table II- Results of document classification accuracy

Type of test data	Classifier training data	Type of features	Classification accuracy
reference	reference	all vocabulary words	77.1% (118/153)
reference	reference	words + classes as keywords	81.7% (125/153)
reference	reference	words + chunks	86.9% (133/153)
ASR	reference	all vocabulary words	67.9% (104/153)
ASR	reference	words + classes as keywords	73.2% (112/153)
ASR	reference	words + chunks	77.8% (119/153)
ASR	ASR	all vocabulary words	69.3% (106/153)
ASR	ASR	words + classes as keywords	70.6% (108/153)
ASR	ASR	words + chunks	74.5% (114/153)
ASR	reference + ASR	all vocabulary words	67.3% (103/153)
ASR	reference + ASR	words + classes as keywords	72.5% (111/153)
ASR	reference + ASR	words + chunks	81.7% (125/153)

In spite of the high WER, very good classification accuracies were obtained by introducing chunks. The main reason for this result is that chunks included classes of words obtained by lists of words such as station and location names, lists of types of documents and objects provided by the call centre service. Words of these classes were frequently used and repeated in the conversations.

By using the value of $\cosine(\bar{X}_d, \bar{X}_m)$ for classification as a confidence indicator it is possible to analyze the trade-off between rejection rate and classification accuracy with the development set. The results obtained in this way are reported in Table III with the test set using manual transcriptions (reference) and ASR results of the train set for training the classifiers and ASR results for test. Precision, recall and F-measure as defined in IR are also reported. A high increase in precision can be obtained with a very high acceptance rate by just rejecting 7.2% of the data. Rejection implies in practice that 11 conversations are transferred to a human expert for analysis. With such a type of classification and precision values, it is possible to automatically infer trends on certain user problems with sufficient accuracy as a larger number of conversations can be made available for test using an approach described in [1].

5. Conclusions

A method for extracting chunks of sentences from ASR transcriptions has been presented. The use of chunks combined with other word features shows significant improvements in a task of conversation classification. The research will continue in the effort of selecting a minimum set of chunks and improving their automatic hypothesization.

6. Acknowledgements

This work is part of the DECODA project supported by the

French National Research Agency (ANR) under contract ANR-09-CORD-005.

7. References

- [1] Camelin, N. De Mori, R. Bechet, F. and Damnati, "Error correction of proportions in spoken opinion surveys", Proceedings of Interspeech, Brighton UK, 2009.
- [2] Chen, Y. N., Huang, Y., Kong Y. K. and Lee, L. S., "Automatic Key Term Extraction From Spoken Course Lectures Using Branching Entropy And Prosodic/Semantic Features", Proc SLT 2010, Berkeley, CA, pp. 253-258.
- [3] Kawahara, T., "New Perspectives on Spoken Language Understanding: Does Machine Need to Fully Understand Speech?", Proc. Automatic Speech Recognition and Understanding (ASRU) workshop, Merano, Italy, dec 2009, pp. 46-50.
- [4] Lavalley R, Clavel C., Bellot P., El-Beze M. "Combining text categorization and dialog modeling for speaker role identification on call center conversations Proc. of Interspeech, Makuhari, Japan 2010
- [5] Linares, G., Nocera, P., Massonie, D. and Matrouf, D. "The LIA speech recognition system: from 10xRT to 1xRT", International Conference on Speech, Text and Dialogue, Pilsen, Tcheck Republic, 2007, Lecture Notes in Computer Science, volume 4629/2007, pp. 302-308.
- [6] Melamed, D. and Gilbert, M. "Speech Analytics".Ch. 14 of [9].
- [7] Oba, T., Hori, T. and Nakamura, A., "Dependency Modeling for Integrated Spontaneous Speech Processing", Proceedings of Interspeech, Lisboa, Portugal, 2005.
- [8] Schapire, R. E. and Inger Y., "Boostexter: a boosting-based system for text categorization," Machine Learning, vol. 39, pp. 135-168, May 2000.
- [9] Tur, G. and De Mori, R. Eds., "Spoken Language Understanding", J. Wiley, March 2011.
- [10] Tur, G. and Hakkani-Tür, D., "Human/Human Conversation Understanding", Ch 9 of [9].

Table III- Results after conversation rejection based on classification confidence

Threshold on $\cosine(\bar{X}_d, \bar{X}_m)$	Classification accuracy	precision	recall	F-measure	Rejection rate
0	125/153	0.81	0.81	0.81	0% 0/153
0,75	122/142	0.86	0.8	0.83	7.2% 11/153
0,8	120/134	0.89	0.78	0.84	12.4% 19/153
0,85	108/120	0.9	0.71	0.79	15.0% 23/153