

Speech Mining in Noisy Audio Message Corpus

Nathalie Camelin¹, Frédéric Béchet¹, Géraldine Damnati²,
Renato De Mori¹

¹ LIA - University of Avignon, BP1228 84911 Avignon cedex 09, France, Fax: 33 4 90 84 35 01

² France Télécom R&D - 2 av. Pierre Marzin 22307 Lannion Cedex 07, France, Fax: 33 2 96 05 35 30

{frederic.bechet,nathalie.camelin,renato.demori}@univ-avignon.fr
geraldine.damnati@rd.francetelecom.com

Abstract

Within the framework of automatic analysis of spoken telephone surveys we propose a robust Speech Mining strategy that selects, from a large database of spoken messages, the ones likely to be correctly processed by the Automatic Speech Recognition and Classification processes. The problem considered in this paper is the analysis of messages uttered by the users of a telephone service in response to a recorded message that asks if a problem they had was satisfactorily solved. Very often in these cases, subjective information is combined with factual information. The purpose of this type of analysis is the extraction of the distribution of users opinions. Therefore it is very important to check the representativeness of the subset of messages kept by the rejection strategies. Several measures, based on the Kullback-Leibler divergence, are proposed in order to evaluate the correctness of the information extracted as well as its representativeness.

Index Terms: Automatic Speech Recognition, Speech Understanding, Confidence Measures, Speech Mining.

1. Introduction

This paper proposes a robust Information Extraction strategy dedicated to process large databases of audio messages. This *speech mining* strategy is particularly dedicated to process *difficult* audio messages like those that can be found in human-human conversations collected in call-centers or users messages left on an answering machine as presented in this study.

Despite all the progress made by Automatic Speech Recognition (ASR) systems, high Word-Error-Rate transcripts are often obtained on speech documents containing bad audio conditions and unconstrained spontaneous speech collected in *real* conditions, as pointed out by the recent NIST Rich Transcription Meeting program or the results obtained on the MALACH corpora. Call centers recordings and telephone survey corpora contain a large variety of speakers, bad audio quality due to cell phones and/or surrounding noises, unconstrained speech, variable utterance length and numerous disfluences like hesitations, repetitions and corrections. As a consequence, speech mining is a very difficult task on such corpora. However the potential applications of speech mining in this context are important: extraction of business intelligence from call center recording or extraction of opinions from telephone surveys.

As far as themes or topics detection is concerned, redundancy of ideas or occurrences of different words directly related to topics can limit the impact of ASR errors. On the contrary, the performance of fine-grained entity detection processes is greatly affected by high WER values. In this case, it is crucial

to select only reliable utterances according to our confidence measures and reject unreliable ones. It is important to notice that rejecting utterances corresponds to extract a sample of the corpus. The performance of an information retrieval system is often evaluated by means of precision and recall measures. The recall measure is not very relevant in our framework as we acknowledge the fact that a possibly significant portion of the data to process has to be rejected because of ASR errors. Therefore we are going to focus on the two following measures: the precision in the information extraction process and an evaluation of the *representativeness* of the sample selected. Indeed it is very important to ensure this representativeness in order to evaluate the bias introduced by sampling.

This paper describes such a sampling and information extraction strategy on the domain of the automatic analysis of spoken telephone surveys. The goal is to extract opinions expressed by callers on several dimensions related to a telephone service. As the average WER obtained on such messages can be as high as 50%, methods have been developed for selecting speech segments in which opinions are reliably detected and reflect the overall distribution of the different opinions on the whole survey corpus. Furthermore, confidence indicators were developed to reject messages considered as not suitable for an automatic process.

Section 2 describes the application used for testing the proposed method for spoken survey analysis. Section 3 describes the system for extracting speech segments in which opinion hypotheses can be extracted with a good level of confidence. Section 3 describes also the rejection method for unreliable messages. Section 4 introduces the method for evaluating strategy results. In this section, experimental results are also provided.

2. Survey Description

2.1. Corpus

The survey corpus used in this study is described in detail in [2] and briefly outlined here.

It has been collected from real users of France Telecom. They are invited through a short message to call a toll-free number where they can express their satisfaction with regards to the customer service they recently called.

During 3 months, about 1.8k messages, with a duration limitation of 2 minutes, were so collected. These messages were transcribed manually and annotated by operators according to the following topics:

1. the courtesy of the customer service operators (*Courtesy*)
2. the efficiency of the customer service (*Efficiency*)

- the amount of time they had to wait on the phone before reaching an operator (*Rapidity*)

Each topic is associate with a positive or negative polarity.

An example (translated from French to English) of a message with its manual segmentation is given below:

yes uh uh here is XX XX on the phone well I've called the customer service yep <courtesy+> the people were very nice </courtesy+> <efficiency+> I've been given valuable information </efficiency+> but <efficiency-> it still doesn't work </efficiency-> so I still don't know if I did something wrong or [...]

Let us call *support* (sup) for the label `courtesy+` the segment *the people were very nice*. We can tag this message with the following labels:

$m : \{sup_1 = \text{courtesy+}, sup_2 = \text{efficiency+}, sup_3 = \text{efficiency-}\}$.

Precision and recall measures will be computed on these labels.

A global dimension was also annotated, experiences were performed and presented in [2].

This corpus was split in two sets : 80% of the messages for the train set and 20% for the test set.

2.2. Analyses

In this study, opinion survey analysis is going to be defined as follows:

Let C be a corpus of n oral messages m_1, m_2, \dots, m_n expressing opinions about a service. Let $C' \in C$ a subset of n' messages of C selected by an automatic analysis strategy.

Opinions expressed in the message are going to be classified according to the topic x and its value v . The value v of a topic x in a message m is defined as follows:

$$v(m, x) = \begin{cases} \text{none} & \text{if } \forall sup_i \in m : sup_i \neq x \\ \text{positive} & \text{if } \forall sup_i \in m : sup_i = x+ \\ \text{negative} & \text{if } \forall sup_i \in m : sup_i = x- \\ \text{mixed} & \text{otherwise} \end{cases}$$

In the following, manual annotations are going to be referred as *ref* (for reference annotations) and the automatic ones given by the strategy developed in this study will be referred as *hyp* (for hypothesis annotations).

The main purpose of opinion analysis is the computation of message proportions containing opinions $O(x, v)$. $O_{ref}(x, v)$ correspond to the manual opinion given to the message m . $O_{hyp}(x, v)$ is the automatic opinion given by the opinion detection module.

The proportion $p_{ref}(x, v)$ given by the manual annotation according to topic x and value v is defined as :

$$p_{ref}(x, v) = \frac{|C(x, v)|}{|C|} \quad (1)$$

with $|C(x, v)|$ corresponded to the number of message $m \in C$ having $O_{ref}(x, v) = TRUE$.

With the interpretations proposed by the system, proportions of hypotheses $p_{hyp}(x, v)$ are defined as:

$$p_{hyp}(x, v) = \frac{|C'(x, v)|}{|C'|} \quad (2)$$

with $|C'(x, v)|$ corresponded to the number of message $m \in C'$ having $O_{hyp}(x, v) = TRUE$.

Let RP be the probability distribution of the reference proportions $p_{ref}(x, v)$ over the different opinions and HP be the probability distribution of proportions $p_{hyp}(x, v)$. Strategies can be compared based on the divergence between the distribution HP they generate and the true distribution RP . The divergence is evaluated by averaging over topics x the Kullback-Leibler divergences (D_{KL}) between the two distributions:

$$D_{KL}(RP(x)||HP(x)) = \sum_v p_{ref}(x, v) \cdot \log \frac{p_{ref}(x, v)}{p_{hyp}(x, v)} \quad (3)$$

The average D_{KL} is defined by :

$$D_{KL}(RP||HP) = \sum_x \gamma_x D_{KL}(RP(x)||HP(x)) \quad (4)$$

where γ_x is a weight proportional to the entropy of the topic x in corpus C .

3. Opinion Detection System

Processing real field telephone data (spontaneous speech of real users in real conditions) is a very difficult task. Users expressed their opinions in a large variety of ways. In such a case, there is often a big mismatch between train and test data and data sparseness is inevitable for modeling all the possible expressions. This results in a high average WER of about 58% on the test set.

The purpose of our system is to detect opinions in users discourse. Typically, users express not only their opinions but also the problem they had, their private situation, they make out-of-domain comments, As a consequence, factual and subjective information is combined.

To take these problems into account, we propose a system based on two steps: transcription and classification. Each step is followed by a rejection process in order to keep for further processes only reliable messages.

A third step introduces use of prior knowledge in the classification and rejection strategy. Different strategies using this prior knowledge are compared in the experiment section.

3.1. Automatic transcription

Instead of a standard word bigram model in charge of transcribing all the words of a given message, ASR models containing specific language models (LMs) dedicated to process opinions are proposed. The purpose of these LMs is to detect message chunks which may convey opinion expressions. These models are presented in [2] and are briefly described here.

A specific ASR decoder is trained on the manually annotated corpus to spot each *support* of topics. A sub-corpus is extracted for each label made of a topic x its polarity $+$, $-$. It contains all the segments associated to this label in the initial training corpus. A specific bigram language model is then estimated on each sub-corpus. Along with these sub-models a global bigram language model is estimated over the 6 labels (3 topics and 2 polarities). A garbage symbol, consisting of an unconstrained contextual phoneme loop, is also estimate to characterize portions of a message which do not express any opinion. The output of this system is a string of segments separated by *garbage* symbols. To each segment is attached the

label of the sub-LM used and the ASR confidence scores.

A classifier is trained in order to filter the segments output according to a set of confidence measures. Segments are represented by their sequence of words, length and ASR confidence scores (acoustic and linguistic). If the score obtained is bigger than a threshold α , the segment is kept for the opinion detection process, otherwise it is rejected.

The main advantage of this ASR system is to directly spot the segments that are likely to contain the expression of an opinion. WER is similar to the one obtain with a standard word bigram model. However, by filtering segments according to confidence measures we can decrease the WER. Let's point out that the method proposed could be used in conjunction with other speech segmentation methods, as the one presented in [6]. However, the main difficulty in this corpus is the nature of the speech: message left on an answering machine with an open prompt; therefore these messages are very hard to structure.

3.2. Opinion Classification

Once the segmentation is done, a classification model, based on a boosting method of *weak* classifiers is applied to the segments selected. It labels each segment with a pair (x, s) where x is a topic and s its polarity sign (*positive* or *negative*). This model is trained with the AdaBoost algorithm [5] on the automatic transcriptions of segments obtained on the training corpus with the ASR opinion specific language models. The opinion labels given to each segment detected are extracted from the reference annotations.

Each segment is represented by three levels of descriptions: POS tags, lemma and *seed* words and its confidence probability. As proposed by other studies [4, 7], a set of words (called *seeds*) which explicitly express a degree of affectedness are manually identified (e.g. *nice, rude, useless, efficient*) and added into the features given to the classifier. A study on the performance obtained with different feature sets is presented in [2].

The real-valued predictions given by the classifier to each pair (x, s) on a segment can be converted into probabilities by passing them through a logistic function [3]. A rejection threshold β is then applied to this probability for filtering the annotated segments: every segment that is classified with a pair (x, s) and a probability above β is kept.

3.3. Use of prior knowledge

The system described below has been implemented and evaluated in [1]. In order to improve performance of this system, a manual study of some classification errors found in the training data is performed. As a result, idiomatic sentences frequently used but not properly taken into account by automatic learning processes are extracted. In our case, this phenomenon is due to the limited size of the training corpus. A limited number of generic, non ambiguous and application-independent expressions are kept. These expressions are generalized, represented by regular expressions and associated with the topic and polarity they are related to.

Let us call this set of patterns *prior knowledge (PK)*. PK is composed of eight patterns for *Courtesy*, two for *Rapidity* and thirteen for *Efficiency*.

Different strategies for integrating PK in our system have been considered. The following strategies have been implemented and evaluated:

- Strategy Ψ_1 is the baseline system without PK.

- Strategy Ψ_2 adds PK features to the input feature set of the *boosting* classifier in Ψ_1 system.
- Strategy Ψ_3 merges opinion hypotheses from the Ψ_1 and opinion hypotheses generated with PK expressions.
- Strategy Ψ_4 applies prior knowledge only to the messages rejected by Ψ_1 and adds these additional opinion hypotheses to the ones generated by Ψ_1 .

All strategies reject some messages. A message is rejected if no topic is detected by the strategy.

4. Experiments

The purpose of the experiments described in this session is to evaluate the different strategies. A standard method for evaluating Information Extraction strategies is the computation of precision and recall measures. For each strategy, precision and recall are evaluated by varying thresholds α and β . The computation of the F-measure, which is a combination of both measures, allows us to choose directly the best strategy considering these standard measures. Results are shown in Figure 1 in which the F-measure is plotted versus the detection precision. Best F-measures are always obtained with the strategies Ψ_3 and Ψ_4 which use prior knowledge.

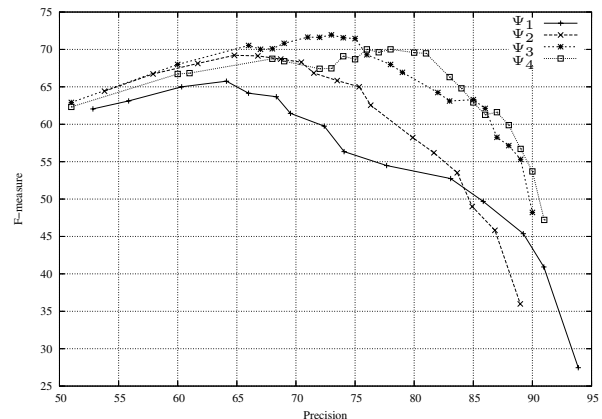


Figure 1: F-measure vs. precision in the opinion extraction on the test corpus for different values of α and β . Curves are obtained on the four different strategies.

As the purpose of opinion extraction is the computation of proportions, strategies have to be evaluated based on the divergence between true proportions and estimated ones. Divergence is computed with the equation 4 for all strategies Ψ_j and different values of α and β . Figure 2 shows D_{KL} as function of precision. Strategies Ψ_3 and Ψ_4 appear to systematically have lower divergence than Ψ_1 .

Low values of precision correspond to low rejection rates. In this situation, the divergence between true and estimated opinion distributions is mostly due to ASR and interpretation errors. High values of precision correspond to high rejection rates. In this situation, high divergence between the distributions is due to the fact that the samples kept are a small portion

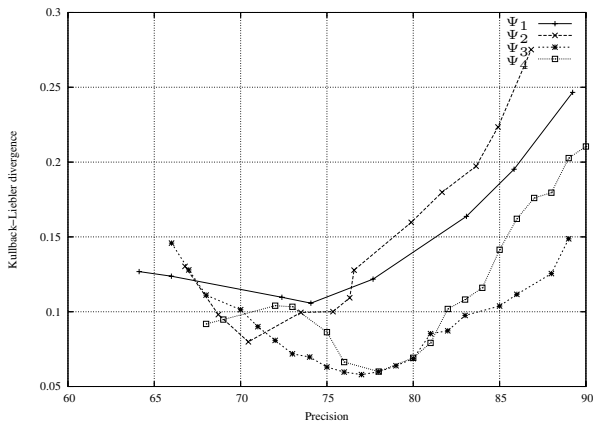


Figure 2: KL divergence between the *true* opinion distribution in the test corpus and the distribution automatically obtained for different values of g leading to different precisions. Curves are obtained for the four different strategies.

of the survey. This subset contains a bias toward one or several opinion dimensions. The lowest divergence is obtained for a precision of about 76% for the strategies Ψ_3 and Ψ_4 . This low divergence value indicates that the filtering process efficiently rejects the problematic messages from an ASR point of view.

As all the strategies include rejection, proportions are estimated with a sample of the survey. In principle, an estimated proportion is affected by two types of errors, the first one is due to ASR and interpretation errors, the second one is due to sampling. If the precision measure is the indicator for the first type of errors, the second type of errors that represent the *representativeness* of the sample selected is not represented by the previous measures introduced.

In order to evaluate this sampling error, we estimate the KL divergence between the true probability distribution of the whole corpus and the true distribution probability distribution of the sample corpus, $D_{KL}(O_{ref})$. Results are plotted in figure 3 for the strategies Ψ_1 and Ψ_4 .

Figure 3 indicates that the sampling process of strategy Ψ_4 does not introduce a bias in the opinion distribution of the messages kept w.r.t the one observed in the whole corpus, even for high precision values like 85%. On the other side, strategy Ψ_1 introduces a bias with precision values bigger than 75%.

5. Conclusion

This paper proposes a method for the automatic analysis of telephone surveys. This system works on very noisy automatic transcriptions of spoken messages and the performance obtained shows the robustness of the method proposed. We introduced several evaluation measures that can be used in order to check the representativeness of the results obtained as well as their accuracy. This point is particularly important in the context of robust information extraction from noisy speech as a selection process has to be used in order to discard the messages that can be processed by the ASR and classification models.

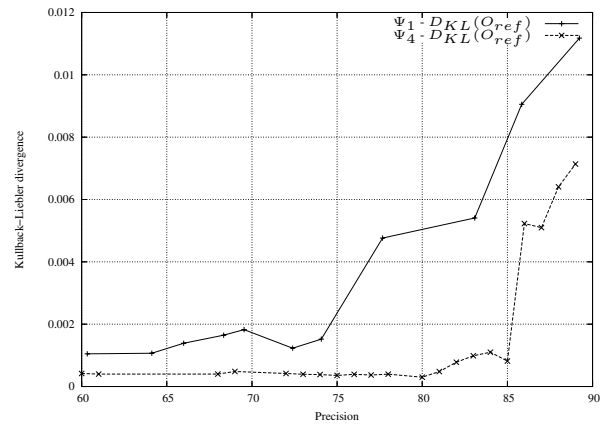


Figure 3: KL divergence between the true probability distribution of the whole corpus and the true distribution probability distribution of the sample corpus obtained for different values of α and β leading to different precisions.

6. References

- [1] Frédéric Béchet, Géraldine Damnati, Nathalie Camelin, and Renato De Mori. Spoken opinion extraction detecting variations in user satisfaction. In *IEEE/ACL Workshop on Spoken Language Technology*, December 2006.
- [2] Nathalie Camelin, Géraldine Damnati, Frédéric Béchet, and Renato De Mori. Opinion mining in a telephone survey corpus. In *Proceedings of the International Conference on Spoken Language Processing*, Pittsburg, USA, September 2006.
- [3] Robert E. Schapire, Marie Rochery, Mazin Rahim, and Narendra Gupta. Boosting with prior knowledge for call classification. *IEEE*, 13(1):174–181, march 2005.
- [4] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of ACL*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.
- [5] Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.
- [6] Gokhan Tur, Andreas Stolcke, Dilek Hakkani-Tur, and Elizabeth Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. In *Computational Linguistics*, volume 27-1, pages 31–57. MIT Press, March 2001.
- [7] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*, pages 347–354, Vancouver, Canada, 2005.