Opinion mining in a telephone survey corpus

Nathalie Camelin[1]*, Géraldine Damnati*[2]*,*
Frédéric Béchet[1]*, Renato De Mori*[1]

[1] LIA - University of Avignon, BP1228 84911 Avignon cedex 09 France
[2] France Télécom R&D - TECH/SSTP/RVA 2 av. Pierre Marzin 22307 Lannion Cedex 07, France
{*nathalie.camelin,frederic.bechet,renato.demori*}*@univ-avignon.fr*
geraldine.damnati@francetelecom.com

Abstract

Telephone surveys are often used by Customer Services to evaluate their clients' satisfaction and to improve their services. Large amounts of data are collected to observe the evolution of customers' opinions. Within this context, the automatization of the process of these databases becomes a crucial issue. This paper addresses the automatic analysis of audio messages where customers are asked to give their opinion over several dimensions about a Customer Service. Interpretation methods that integrate automatically and manually acquired knowledge are proposed. Experimental results, done on a database collected from a deployed Customer Service in real conditions with real customers are given.

Index Terms : speech recognition, language models, opinion mining.

1. Introduction

Over the past several years, there has been an increasing number of publications focused on opinion mining or sentiment and subjectivity detection in text [2]. Two main application domains are targeted by such research :

- the automatic analysis of surveys : in business intelligence it is important to automatically extract positive and negative perceptions about features of a product or service [5] ;
- Information Extraction, summarization, and question answering applications : in these applications it is important to distinguish highly subjective stance from a mostly objective presentation of facts [6].

An emerging application of telephone services consists in asking opinions from the users about the solution of a problem for which they have called a special number. Users typically describe their problem and express opinions on the way it was treated. Users often describe their problem with more than one sentence and a certain amount of redundancy. Furthermore, the type of problem and the level of user satisfaction are only part of the semantic content of the user discourse. Opinion mining constitutes a problem that is orthogonal to typical topic detection tasks in message classification. A user can be totally or partially satisfied for part of a service and not satisfied for other aspects. User satisfaction analysis has complex dimensions which include, but are not limited to automatic Spoken Language Understanding (SLU). User messages may contain a variable number of sentences of highly variable length. Repetitions, hesitations, reformulation of the same concept are frequent. All types of redundancy are useful to recover from Automatic Speech Recognition (ASR) errors.

One of the main originalities of this work is to process spoken telephone surveys collected from *real* users. Therefore we deal here with all the problems linked to the processing of very spontaneous speech (telephone speech, real users, bad audio quality due to cell phones and surrounding noises, unconstrained speech). Section 2 presents the corpus used in this study. The Automatic Speech Recognition (ASR) process adapted to this specific task is presented in section 3, the classification methods in section 4.1. An evaluation of the robustness of the proposed methods is given in section 5.

2. A corpus of telephone surveys

Users 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. Calling this toll-free number, they are asked to leave a message, after hearing the following prompt : "*[...] You recently contacted our customer service. Thank you for calling us. We would like to make sure that you've been satisfied with the reception and with the way your request was followed up. Don't hesitate to make any comment or suggestion about our service that can help us make it better.[...]*"

These messages are currently processed by operators who listen and qualify the messages according to a variety of criteria for further statistical analysis. A subset of 1779 messages, collected over a 3 months period has been additionally transcribed manually in order to train models to perform automatic message qualification.

Two kinds of opinion expression have been manually annotated : one related to the global satisfaction (called *SatGlob* in this study), represented by an indication of the perceived user judgment on the service (positive, negative or neutral) ; the other one is a finer grained opinion analysis where four dimensions are used in order to characterize users opinion :

1. the courtesy of the customer service operators (*Courtesy*)
2. the efficiency of the customer service (*Efficiency*)
3. the amount of time one has to wait on the phone before reaching an operator (*Rapidity*)
4. the last dimension groups together all the expressions of opinion on other subjects that the first three ones (*Other*)

These last four criteria can receive two polarities : positive or negative, leading to a set of 8 opinion labels. In the manual transcription of this corpus each message is labeled with its global sa-

collaboration with France Telecom R&D - contract N 021B178

tisfaction label (*SatGlob*= {*positive,negative,neutral*}). Within a message, each opinion expression (positive or negative) on the four dimensions (*Courtesy*, *Efficiency*, *Rapidity*, *Other*) is segmented thanks to markup tokens. The goal of the opinion mining process is to automatically retrieve these opinion labels.

Annotating manually opinion expressions can be a difficult task for some ambiguous messages. This is particularly the case for semantic labels that can't be precisely defined, like the global satisfaction (*SatGlob*) label. Therefore it is important to verify the consistency of the manual annotations by means of several annotators each processing the same set of messages in order to measure their agreement [1]. The Kappa (K) measure is accepted as a reliable inter-annotator measure. A value of 0.7 or above is generally considered as a correct agreement value. For the *SatGlob* label, the Kappa measure has been estimated on a set of 70 messages with four annotators. A value of $K = 0.6$ has been obtained with the 3 labels *positive*, *negative* and *neutral*. By considering only the messages labeled *positive* or *negative* (all the messages labeled *neutral* by at least one annotator are discarded), we obtain a Kappa value of $K = 0.9$ and these messages represent 90% of the corpus. This means that it is the *neutral* label that is the most ambiguous, but there is very little ambiguity between *positive* and *negative* labels.

# opinion	% corpus	message avg. length
0	19.2	61.0
1	51.3	40.3
2 and more	29.5	60.8

TAB. 1 – Distribution of the messages in the corpus, with the average message length, according to the number of opinion expressed.

The average length (in words) of a message according to the number of opinion expressed is presented in table 1. Even if the messages expressing only one opinion are the shortest, the length of a message in itself is not a reliable indicator as the longest messages in average are those expressing no opinions on the four dimensions already mentioned. This does not mean that these messages don't express a global satisfaction sentiment. Another difficulty of this kind of corpus is that a single message can contain several times the same opinion expression with different polarities. This happens when people have mixed feeling about the service, or when they refer to several customer experience they had. 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* <other-> *it still doesn't work* </other-> *so I still don't know if I did something wrong or [. . .]*

Messages are recorded in totality with a duration limitation of 2 minutes. After processing a noise/speech detection to cut the initial and final silences, continuous speech recognition is performed on messages. As a consequence to the recording conditions (a message left on an answering machine) the language is highly disfluent. 30 % of the messages contain at least one truncated word as a result of a false-start. The average number of filled pauses per message is 3.7.

3. Opinion specific language models and automatic segmentation

3.1. Baseline ASR model

Due to the lack of constraints on users' elocution and to the nature of the open question they are submitted to, a large dispersion can be observed in the word frequency distribution. This phenomenon is particularly observed in those portions of messages where users recall the origin of their problem (which is usually fairly different from a user to another). Once Named Entities have been parsed (such as phone numbers, last names..) and replaced by a single label, the training corpus contains close to 3000 different words for a total of 51k occurrences. Nearly half of these words occur just once and the restriction to those words that occur at least twice led to a lexicon of 1564 words, for a 2.8% out-of-vocabulary rate. A first bigram language model has been estimated with this lexicon. Because of the very high level of disfluencies and noise, especially in long messages, the WER obtained with this model is high : 58% on average. However the WER is not the same for all messages, for example short messages obtain better performance, as shown in table 2. As a matter of fact longer messages contain more digressions with a higher OOV rate.

WER	<20	<30	<40	<50	<60	> 60
length (words)	17.7	22.6	36.2	51.9	65.0	71.3

TAB. 2 – Correlation between WER and message length (in words)

3.2. Automatic opinion segmentation

In parallel, a first attempt towards an automatic segmentation of messages has been achieved. The objective is to facilitate the classification task by providing fragments of messages instead of the whole (potentially long) message. This first attempt consists in segmenting messages by means of acoustic features, through the detection of pauses. Even if there is no *a priori* correlation between the presence of a pause and a thematic change, this first approach will be used as a baseline for the rest of the study. Segments of signal that are isolated are submitted independently to the speech recognition system and the corresponding recognition hypotheses are transmitted to the classification modules. These outputs will be referred to as *RECO1* in section 5.

This segmentation turned out to be insufficient. As a matter of fact, long segments carrying several opinion expressions (pronounced without any pause between them) are still remaining while some segmentation are made in the middle of a single opinion expression. In a second approach, the problems of segmentation and recognition have been integrated through a new type of language model. The idea is to explicitly model only those portions of messages that carry opinion expressions. To this end, a sub-corpus has been extracted for each opinion label, containing all segments associated to this label in the initial training corpus. A specific bigram language model has then been estimated on each sub-corpus. Along with these sub-models a global bigram language model has been estimated over a label lexicon of size 9 (the 8 opinion labels themselves and a garbage label modeling portions that do not correspond to any opinion expression). This global LM enables to model the possible coocurrences of opinions in a single message.

In order to obtain a single fully compiled recognition model, each occurrence of an opinion label in the global LM is replaced by the corresponding sub-LM. The garbage model consists of an unconstrained contextual phoneme loop. This unique, fully compiled model is referred to as *RECO2* in section 5.

On the overall, 1117 segments have been extracted for all the opinion labels, corresponding to 18700 occurrences of words. The number of different words per sub-corpus is not higher than 780, with an average of 470. The first interest is therefore to have largely reduced the lexical field. From another point of view, messages are globally characterized by a high disfluency level. But then again, the most disfluent portions are not the ones where users express their opinion, but mostly the ones where they recall their initial problem. We can then observe a reduction of the disfluency level within the extracted opinion segments. This is illustrated in table 3.

Indicator	# messages	# segments
filled pauses	6.1	5.0
false starts	1.9	1.7
restarts	4.2	3.9
repetitions	2.0	2.3
discourse markers	4.3	1.2

TAB. 3 – Percentage of disfluency indicators in the initial corpus and in the extracted corpus

Besides repetitions that do not constitute the most problematic phenomena from the recognition point of view, all the indicators are lower in the extracted opinion segments. The most important decrease is observed for the discourse markers which are difficult to model (due to the variety of potential contexts) and especially whose ambiguities can disturb the a posteriori treatments on the messages. For example the words *bon* or *bien* (both corresponding to *well*) can be either meaningful for an opinion or neutral when they are used as discourse markers.

The two segmentation process proposed need now to be integrated in order to improve the relevance of the segments, following previous works like [8].

4. Modeling user satisfaction

Analyzing user satisfaction from message transcriptions is a special Information Extraction task that combines fine grained entity extraction for the detection of a user's opinion on a particular dimension and thematic classification when categorizing message contents. Section 4.1 shows how some classification methods previously used for call routing or message classification can be used for this purpose and section 4.3 presents how message segmentation models are needed in this framework.

4.1. Classification method

Previous works on message classification or call routing [3] have used classification methods, like Boosting [7] or Support Vector Machines [4], for labeling an utterance with one or several labels (called *calltypes* in this study) corresponding to the global meaning of the utterance. Different classification algorithms and data representations of an utterance (bag of words, n-grams, ...) are used. The information extraction task presented in this study is related to these previous works as some of the labels to be detected are directly related to the theme of a message (like *Courtesy*) and therefore can be considered as calltypes. However, as themes and judgments may be integrated into linguistic structures, the task is more complex and directly applying techniques developed for extracting calltypes to user satisfaction analysis has some limitations.

The classification tool used in this study is *BoosTexter* [7]. This is a large-margin classifier based on a boosting method of *weak* classifiers. The weak classifiers are given as input. They can be the occurrence or the absence of a specific word or n-gram, a numerical value (like the utterance length) or a combination of them. At the end of the training process, the list of the selected classifiers is obtained as well as the weight of each of them in the calculation of the classification score for each conceptual constituent of the tagset.

4.2. Using prior knowledge

Previous studies on sentiment analysis in text have shown the usefulness of integrating prior knowledge in the classification process by means of lexicons containing words that are likely to be associated with the expression of an opinion. These words, called *seeds* words, are likely to attach a positive or negative polarity to a stance [9]. To this purpose, a set of words that explicitly express a degree of affectedness are then identified (e.g. *satisfied*, *rude*, *problem*, *fixed*).

When a labeled training corpus is available, supervised learning methods can be used to automatically infer *seed* words, related to the application. The following method is used : the training corpus, made of the exact word transcriptions of the messages as well as the reference user opinion labels, is used to train the text classifier BoosTexter ; after n iterations, n weak classifiers are obtained, each of them representing the occurrence of a word or an n-gram sequence of words. All the words selected are then added to the manual lexicon of seed words. The last step in this process consists in replacing each word by its lemma for augmenting the generalization capabilities of the classification model.

With this method a set of 565 lemma have been selected. A message can be represented by all its words (manual or automatic transcriptions) or by the seed words contained in it. The features chosen in our experiments are 1-gram, 2-gram and 3-gram features on the words or the seed words.

4.3. Segmenting messages according to opinion expressions

As seen in section 2, several opinion expressions can occur in a message. If for the global satisfaction label the messages can be given as a whole to the classification process, a segmentation process is needed for the four opinion dimensions (*Courtesy*, *Efficiency*, *Rapidity*, *Other*). For the reference transcriptions of the corpus, this segmentation in opinion is manually given. For the automatic transcriptions, two methods are compared : a baseline one segmenting a message based on pauses in the speech signal (called $RECO_1$ in the experiments) and the one presented in section 3 ($RECO_2$).

5. Experiments

The classification models are trained on the same training corpus as the one used in section 3 for the language models. The test corpus, containing 580 messages, is processed according to 3 conditions :

- *REF* : the reference transcriptions (with manual opinion segmentation) ;
- $RECO_1$: automatic transcription, segmentation done ac-

cording to pauses in the speech signal ;

– $RECO_2$: for the 8 opinion labels, transcription with the language model presented in section 3, segmentation done by the ASR process.

For each condition the results are given according to two data representation : messages represented only by their words or filtered according to the seed words. Precision (P), Recall (R) and F-measure are presented for each condition in the following sections.

5.1. Global satisfaction

Because no segmentation is needed (the *SatGlob* label applies to the whole message), only $RECO_1$ is used here. The results are reported in table 4.

polarity	words			seeds		
+-0	P	R	F	P	R	F
REF	72.4	67.0	69.6	73.5	69.2	71.2
$RECO_1$	67.3	62.0	64.5	69.3	63.0	66.0
+-	P	R	F	P	R	F
REF	81.7	81.7	81.7	81.4	81.4	81.4
$RECO_1$	74.3	74.3	74.3	76.3	76.3	76.3

TAB. 4 – Performance on the *SatGlob* opinion label detection, both on manual transcriptions (*REF*) and automatic transcriptions ($RECO_1$). Results are given with 3 polarities (positive, negative, neutral *+-0*) and 2 polarities (+-).

As we can see, the results are much better when considering only the messages containing an explicit opinion (positive or negative). When adding the neutral polarity, the performance is slightly worse, matching the poor Kappa inter-annotator agreement measure obtained with 3 polarities. One very interesting result in this table is the limited impact of dealing with highly erroneous transcripts instead of manual reference transcriptions : the loss in F-measure is only 5% despite the very high WER of the automatic transcriptions. Representing the messages by means of seed words seems also to increase the robustness of the classification.

5.2. Fine grain opinion detection

Table 5 presents the results obtained with the 8 opinion labels (the four dimensions *Courtesy*, *Efficiency*, *Rapidity* and *Other* with two polarities *positive* and *negative*. The two segmentation methods proposed, $RECO_1$ and $RECO_2$, are compared to the reference transcription with the manual segments. Each message is represented by a set of segment. Each segment is processed by the classification module in order to receive an opinion label. The set of opinions attached to a message is the concatenation of the local decision on every segment of the message.

	words			seeds		
(en%)	P	R	F	P	R	F
REF	75.8	60.0	67.0	77.1	64.2	70.0
$RECO_1$	51.6	36.2	42.6	49.5	37.6	42.7
$RECO_2$	41.1	54.7	46.9	42.0	57.0	48.4

TAB. 5 – Comparison of 3 message segmentations for the detection of the 8 opinion labels

Although the segmentation method $RECO_2$ outperforms the baseline one based on pause detection, the results are this time very far from those obtained on the manual transcriptions. This is mainly due to the insertion of segments by $RECO_2$, leading to a good recall value but a poor precision. We are now focusing on the use of confidence measure in order to filter the segments generated by $RECO_2$ in order to increase the precision score. One can also see that the use of seed words increases the performance for every conditions.

6. Conclusions

Interpretation methods that integrate automatically and manually acquired knowledge have been proposed. The experiments carried on with the global satisfaction detection have shown that even very noisy automatic transcripts, with a very high WER, can be efficiently processed. When dealing with more complex opinion expressions, a segmentation module is needed. This work presents a first module, integrated into the Automatic Speech Recognition process, that produce transcriptions only for portions of a message that are likely to be an expression of an opinion on one of the dimensions targeted. Despite an improvement obtained with respect to the baseline segmentation method based on pause detection, the results are still quite far from those obtained with a manual segmentation. Therefore we are now investigating discriminant segmentation models, like those based on Conditional Random Field, in order to integrate them in the decoding process.

7. References

[1] Jean Carletta. Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics*, 22(2) :249–254, 1996.

[2] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhann. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proc. of HLT/EMNLP*, pages 355–362, Vancouver, 2005.

[3] Allen L. Gorin, Giuseppe Riccardi, and Jeremy H. Wright. How may I help you ? *Speech Communication*, 23(1-2), 1997.

[4] Patrick Haffner, Gokhan Tur, and Jerry Wright. Optimizing SVMs for complex call classification. In *Proc. IEEE ICASSP'03*, Hong-Kong, 2003.

[5] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proc. of HLT/EMNLP*, pages 339–346, Vancouver, 2005.

[6] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proc. EMNLP*, 2003.

[7] Robert E. Schapire and Yoram Singer. BoosTexter : A boosting-based system for text categorization. *Machine Learning*, 39 :135–168, 2000.

[8] Gokhan Tur, Dilek Hakkani-Tur, Andrea Stolcke, and Elizabeth Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics - MIT Press*, 27-1 :31–57, March 2001.

[9] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of HLT/EMNLP*, pages 347–354, Vancouver, 2005.