SPOKEN OPINION EXTRACTION FOR DETECTING VARIATIONS IN USER SATISFACTION

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

In recent years, efforts have been made for automatically identifying opinions, emotions and sentiments in text. 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 opinion analysis is the detection of time variations of user satisfaction indices. Even if precision or recall is not very high because messages are ambiguous or ASR systems have made many word recognition errors, system strategies are acceptable if they detect the same trend in user satisfaction as it is indicated by human interpreters of the messages. In this paper a system for this type of opinion analysis is proposed for a telephone service survey task.

Index Terms— Automatic Speech Recognition, Speech Understanding.

1. INTRODUCTION

In recent years, efforts have been made for automatically identifying opinions, emotions and sentiments in text [1, 2]. Sentiment classification is a categorization task in which the goal is to classify a document as having positive or negative polarity. Other research efforts analyze opinion expressions at the sentence level and below to recognize opinions, their polarity and strength.

The result of the process is the generation of a large set of labeled sentences. Extraction patterns are used to represent subjective expressions. Most previous work on subjectivity classification has focused on document-level classification. A variety of algorithms have been developed to automatically learn extraction patterns. Some of them require texts annotated with domain-specific tags, manually defined key-words, frames or object recognizers, separation between relevant and irrelevant texts, bootstrapping with seed words and seed patterns.

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. User satisfaction requires the extraction of subjective information from one or more spoken sentences. A spoken message contains segments which convey opinion information and others which are not relevant for opinion analysis. User satisfaction is expressed by concepts like *efficiency* and their polarity. Such an expression has a *diffuse support* made of words in a spoken message. The diffuse support is not limited to one or more syntactic structures which, in any case, cannot be obtained because Automatic Speech Recognition (ASR) system output contains many errors and spoken language often is not generated by known grammar rules.

Furthermore, detecting opinions with diffuse support prevents to use dynamic programming approaches because supports of different opinions may overlap. A decision should rather be based on the computation of the probability that an opinion hypothesis is present in a spoken message. Subjective information useful for actual telephone services comes from a single source and does not have a complex structure. On the other hand, analyzing spoken messages rather than written text has to deal with errors of ASR systems and the fact that spoken language contains hesitations, repetitions, and corrections, incomplete and ungrammatical sentences. Specific solutions have to be proposed to take these problems into account.

The purpose of this type of opinion analysis is the detection of variations of user satisfaction trends (i.e. when the ratio positive *vs.* negative opinion changes). Even if precision or recall is not very high because messages are ambiguous or ASR systems have made many word recognition errors, system strategies are acceptable if they detect the same trend in user satisfaction as it is indicated by human interpreters of the messages. In this paper a system for this type of opinion analysis is proposed.

2. SYSTEM OUTLINE

Automatic opinion analysis from real field telephone data is a very difficult task. A large variety of unpredictable speakers express their opinions in different ways with spoken messages of highly variable length, with possible repetitions, corrections and contradictions and in a variety of acoustic environments. This results in a large variety of automatic speech recognition (ASR) system performance. A first prototype described in this paper does not solve all the problems. Nevertheless, attention is paid to the conception of a strategy that automatically detects a change in users satisfaction trend.

Let *message m* be the oral document to be analyzed. The system generates hypotheses about services satisfaction and their polarity. Service satisfaction is expressed in terms of *Efficiency*, *Courtesy* and *Rapidity*. Polarity is positive or negative. The possibility that a message does not contain any expression of satisfaction is also considered. Let these seven possibilities be values of a variable indicated as *satexpr*.

Let H be a hypothesis about a *satexpr*. It may contain a service attribute and a polarity. Let W_H be a pattern of word hypotheses

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

generated by an ASR system based on which the hypothesis H has been generated. Pattern W_H is called a support for H. Let $A(W_H)$ be the sequence of vectors of acoustic features based on which the hypothesis W_H has been generated. Notice that different compatible *satexpr* hypotheses may share some words in their supports. The hypothesis H is scored by the following probability:

$$P(H|A) = \sum_{W_H} P(H, W_H|A) \approx \max_{W_H} P(H|W_H)P(W_H|A(W_H)) \quad (1)$$

A represents the acoustic features of the entire message, $P(W_H|A(W_H))$ is computed from the scores provided by the ASR system. $P(H|W_H)$ is computed by the interpretation module.

The interpretation module attempt to perform a segmentation of the spoken message by separating word sequences carrying information about a *satexpr* from other words. Segmentation knowledge is automatically learned by training statistical models using a manually annotated corpus.

3. A CORPUS OF TELEPHONE SURVEYS

The corpus used in this study is presented in [3] and is briefly described here. 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. A study on the global opinion detection of the callers is presented in [3]. In this paper we focus on a finer grained opinion analysis considering three dimensions:

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*)

These criteria can receive two polarities: positive or negative, leading to a set of six opinion labels. At the message level these criteria and polarities are combined in order to predict four opinion expressions on the three dimensions:

- *positive*(+): contains only positive polarities for the dimension considered;
- *negative*(-): contains only negative polarities for the dimension considered;
- *mixed*($\sim$): contains positive and negative polarities for the dimension considered;
- *none*(0): this dimension is not in the message.

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 [. . .]*

This message would receive the following opinion labels: *Courtesy=positive*, *Efficiency=mixed*, *Rapidity=none*.

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, no dialogue) the language is highly disfluent, the signal is often noisy, and the messages are quite long (an average of 50 words).

4. AUTOMATIC TRANSCRIPTION OF OPINION MESSAGES

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. 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 longer messages contain more digressions with a higher OOV rate.

Because of the high WER in the automatic transcriptions produced, it is very important to estimate a confidence score on each word produced. The following types of confidence indicators are proposed:

- AC, a descriptor of acoustic confidence based on the comparison of the acoustic likelihood provided by the speech recognition model for a given hypothesis to the one that would be provided by a totally unconstrained phoneme loop model;
- LC, a descriptor of linguistic confidence that estimates the number of ngrams (in a segment or surrounding a word) that has been observed in the training corpus compared to those that lead to a backoff from the language model.

The probability for a word or a segment of being correct is approximated on the training corpus by means of logistic regression:

$$P(Correct|AC, LC) = \frac{1}{1 + e^{-(a_0 + a_1 \times AC + a_2 \times LC)}}$$

On the training corpus, the weights obtained are: $a_0 = -0.62$, $a_1 = 0.05$ and $a_2 = 2.22$.

5. MESSAGE SEGMENTATION AND OPINION EXTRACTION

According to the model presented in section 2, finding an opinion H in a message M consists in detecting a word support W_H for H in M then estimating the 2 probabilities $P(H|W_H)$ and $P(W_H|A(W_H))$. Considering that the second probability is given by the ASR models, the following subsections present the model used for estimating $P(H|W_H)$ and for extracting W_H from M.

5.1. Opinion classification

The classification tool used in this study is *BoosTexter* [4]. 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.

$P(H|W_H)$ is approximated from the classification scores given by BoosTexter with a logistic regression process applied on a development corpus.

The training corpus of the classifier is made of the manually labeled segments of our telephone survey corpus. To each word of a segment are attached 3 tokens representing 3 levels of description: POS tags, lemma and *seed* words. As proposed by other studies [2], a set of words (called *seeds* that explicitly express a degree of affectedness are identified (e.g. *satisfied*, *rude*, *problem*, *fixed*). Specifying explicitly key words can help the classification process as BoosTexter uses ngram features, therefore ngrams of key words can be used thanks to these seed words.

5.2. Segmentation with Opinion-specific Language Models

A specific ASR decoding model has been defined in order to spot directly the portions of the messages likely to contain the expression of an opinion. This model is presented in [3] and is briefly described here. A sub-corpus is extracted for each opinion label, containing all 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 a label lexicon of size 7 (the 6 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 used by the ASR engine in order to produce the automatic transcriptions. 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.

The last step in this segmentation process is to reject the segments with a low confidence and to merge the consecutive segments sharing the same opinion label (i.e. produced by the same sub-LM). The WER obtained on the segments selected is 52% (compared to 58% with the general LM). In addition to the small gain in WER, the main advantage of these opinion-specific LMs is to directly produce segments that can be used for the opinion extraction process.

Let's point out that the method proposed could be used in conjunction with other speech segmentation methods, as the one presented in [5]. 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.

6. OPTIMIZING OPINION EXTRACTION FOR AUTOMATIC SURVEY ANALYSIS

When performing classification tasks, systems can be optimized in order to maximize the precision of the classification, the recall, or a combination of them through the F-measure. For the user satisfaction survey task presented in this study, the size of the set of messages processed is not relevant as long as the distribution of the opinions over this set is similar to the true distribution in the whole corpus. Therefore the only measure to optimize is the distance between the opinion distribution obtained on the subset of messages selected by the system and the *true* distribution on the whole corpus. This distance can be estimated with the KullbackLeibler (KL) divergence. In the system presented in this paper, 2 parameters needs to be set for defining an interpretation strategy: the threshold α on the probability $P(H|A)$, which decides if the opinion H is accepted for the message M represented by the acoustic features A; the threshold β on the ASR confidence scores that accepts or rejects the segments produced by the opinion-specific LMs.

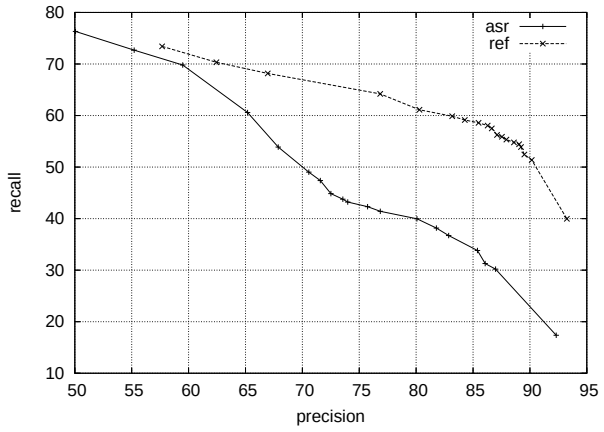

Fig. 1. Precision *vs.* recall in the opinion extraction on the development corpus for different values of α and β. The curve *asr* is obtained on the automatic transcriptions, *ref* on the manual reference transcriptions.

A development corpus is used for tuning these 2 parameters. For each value of α and β several measures are computed: precision and recall on the detection of all the opinion tags; the KL divergence of the distribution of the four opinion expressions: *positive*(+), *negative*(-), *mixed*($\sim$), *none*(0) on the 3 dimensions: *Courtesy*(C), *Rapidity*(R) and *Efficiency*(E). The average KL divergence for these 3 dimensions between the reference distribution P_{ref} on the development corpus and the one hypothesized by our system P_{hyp} is estimated as follows:

$$D_{KL}(P_{ref}||P_{hyp}) = \frac{1}{3} \times \sum_i \left(\sum_j P^i_{ref}(j) \log \frac{P^i_{ref}(j)}{P^i_{hyp}(j)} \right)$$

with $i \in \{C, R, E\}$ and $j \in \{+, -, \sim, 0\}$. The values chosen for α and β are those minimizing $D_{KL}(P_{ref}||P_{hyp})$ on the development corpus.

7. EXPERIMENTS

From the telephone survey corpus presented in section 3, 2 corpora have been extracted: a training corpus containing 1150 messages and a development corpus containing 360 messages. Figure 1 shows the precision and recall measures obtained by varying the parameters α and β on the development corpus. For the reference transcriptions (noted *ref* in this figure) only the parameter α is used and the segmentation process is performed thanks to a Conditional Random Field tagger based on the toolkit *CRF++* [1]. As one can see, for a precision of 80%, there is a drop of about 20% in the recall measure between the reference transcriptions and the automatic ones (from

[1] http://www.chasen.org/ taku/software/CRF++/

60% to 40%). This is not surprising considering the very high WER in the automatic transcriptions. This low recall on the automatic transcriptions is not necessarily problematic if the distribution on the remaining messages is similar to the one on the whole corpus.

Figure 2 shows this measure by plotting the KL divergence between the *true* distribution on the development corpus and the one obtained on the set of messages kept by the interpretation strategy. The KL is given for the same values of α and β as those used in figure 1.

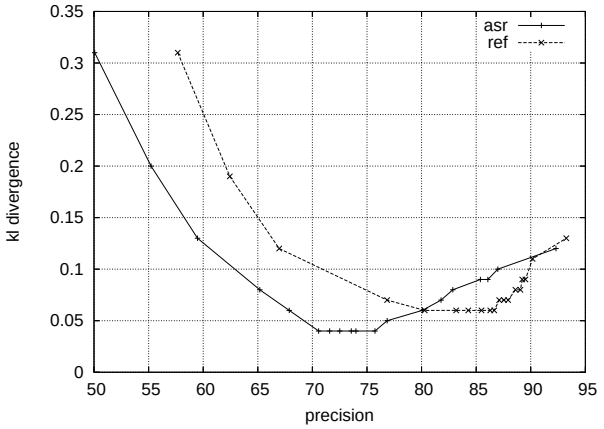

Fig. 2. KL divergence between the *true* distribution on the development corpus and the distribution automatically obtained for different values of α and β. The curve *asr* is obtained on the automatic transcriptions, *ref* on the manual reference transcriptions.

The lowest KL values are obtained for a precision of 80% and a recall of 70% on the manual transcriptions and a precision of 70% for a recall of 50% on the automatic transcriptions. In both cases the KL divergence value is about 0.05 bits. In comparison, a flat distribution over the opinion expressions has a KL divergence value of 0.8 bits. These lowest KL values are the operating points chosen for the opinion survey system.

In order to check the ability of this system in detecting a change in the trend of users satisfaction, several test corpora are selected from the telephone survey corpus. These corpora differs according to the proportion of positive opinions for the dimension *Efficiency* (E), from 30% to 70%. The opinion extraction system presented in this paper is applied on the audio messages of each of these corpora, and the distribution of opinion on the three dimensions (*C, E, R*) is collected. Figure 3 presents the variation in the automatically obtained distributions between the positive and negative opinions for dimension E. These variations are measured from a starting point corresponding to the proportion of positive opinions for E in the training corpus ($x = 54\%$). On the y-axis is plotted the variation (%). As one can see, if the % of positive opinions decreases in the test corpus ($x < 54\%$), the variation in the distribution of negative opinions is > 0 and increases while x decreases. Similarly this variation is < 0 for the positive opinions and the same trend is noticed when $x > 54\%$. We can see that the variations in the test corpora are correctly detected by our system and therefore it can be use as a relevant indicator of a change in users satisfaction trend.

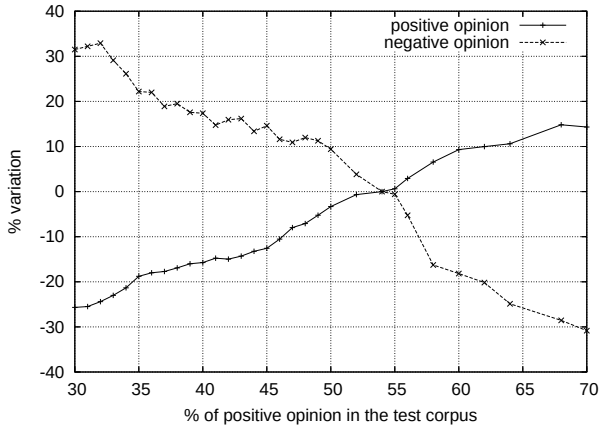

Fig. 3. Variation (%) in the distributions automatically obtained of positive and negative opinions for *Efficiency* with different test corpora containing different proportions of positive opinion. The starting point for measuring the variations is 54% positive opinion (as in the training corpus).

8. CONCLUSION

This paper proposes a method for the automatic analysis of telephone surveys. The model proposed can accurately detect a variation in the users satisfaction trend on a specific dimension. This system works on very noisy automatic transcriptions of spoken messages and the performance obtained shows the robustness of the method proposed. However, because our measures are affected by errors, we are working now on modeling the bias introduced by the recognition errors according to the ASR confidence scores attached to the transcriptions. Estimates of change from one time period to another (or from one set of confidence values to another) remains unbiased provided that the bias is constant throughout.

9. REFERENCES

[1] Jayce Wiebe, Theresa Wilson, and Claire Cardie, "Annotationg expressions of opinions and emotions in language," in *Language Resources and Evaluation*, 2005, vol. 39, pp. 165–210.

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

[3] Nathalie Camelin, Géraldine Damnati, Frédéric Béchet, and Renato De Mori, "Opinion mining in a telephone survey corpus," in *Proc. of Interspeech-ICSLP'06*, 2006.

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

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