INTERSPEECH 2009 BRIGHTON

Error correction of proportions in spoken opinion surveys

Nathalie Camelin[1], *Renato De Mori*[1], *Frederic Bechet*[1], *Geraldine Damnati*[2]

[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

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

10.21437/Interspeech.2009-694

Abstract

The paper analyzes the types of errors encountered in automatic spoken surveys. These errors are different from the ones that appear when surveys are taken by humans because they are caused by the imprecision of an automatic system. Previous studies presented a strategy that consists in the robust detection of subjective opinions about a particular topic in a spoken message. If the same automatic system is used for estimating opinion proportions in different spoken surveys, then the error rate of the entire automatic process should not vary too much in different surveys for each type of opinions. Based on this conjecture, a linear error model is derived and used for error correction. Experimental results obtained with data of a real-world deployed system show significant error reductions obtained in the automatic estimation of proportions in spoken surveys.

Index Terms: Automatic Speech Recognition, Speech Understanding, Opinion analysis, Error correction.

1. Introduction

In recent years, efforts have been made for automatically identifying opinions, emotions and sentiments in text [8, 9] with particular attention to news articles and product reviews [6, 7]. Other research efforts deal with review classification and mining, detection of opinion holders and topic of opinions [1, 6].

Recently, in [2, 3], a system has been described for the automatic processing of spoken surveys. These surveys are made on opinions about telephone services in response to a recorded message asking a user if a previously signaled problem was satisfactorily solved. User responses are often fairly long messages containing subjective and factual information. Subjective information can be expressed in one or more spoken phrases that are the *support* with which an opinion is expressed.

The automatic opinion analysis and survey are performed on telephone data uttered by a large variety of public users in different acoustic environments. Messages from these users are processed by an Automatic Speech Recognition (ASR) system which is error prone. Other difficulties for automatic interpretation arise from the fact that unpredictable speakers express the same opinion in many ways with spoken messages of highly variable length, with possible repetitions, self-corrections and contradictions.

A system[1] described in [3] provides interpretation hypotheses from which opinion proportions are computed. Such a system uses conceptual Language Models (LMs) for detecting message chunks which may contain opinion and subjectivity information. In this way, an initial separation between pertinent and factual speech segments is performed. Classifiers are then used for generating hypotheses about opinions and evaluating their confidence. In order to cope with ASR and interpretation errors, messages that are not automatically transcribed with a sufficient confidence are discarded. Opinion proportions are estimated on the data that are not rejected. Such estimation is affected by two types of errors, one due to rejection and another due to the imprecision of the whole process in interpreting the non rejected messages.

An important problem arising when surveys are taken concerns the measurement errors, the possibility of deriving models for them and of using the models for compensating the effects of errors [4]. The types of errors encountered in automatic spoken surveys are different from the ones that appear when surveys are taken by humans. For example, automatic rejection of messages does not have the same characteristics as usual population sampling techniques and errors of automatic systems are different from the ones affecting the interviews made by humans.

In [3], a method is proposed for tuning some parameters of the opinion detection system in order to minimize the relative entropy between the true and the estimated distributions of opinion proportions. In [5], an algorithm is proposed for estimating biases when proportions of class hypotheses are obtained from the output of classifiers, assuming accurate knowledge of classifier error rates. A method is proposed in this paper that derives an error model by just considering a linear approximation of the relation between estimation errors and computed proportions.

After a formulation of the proportion estimation problem in section 2, the types of spoken survey errors are introduced in section 3, together with a linear model for error correction. Section 4 describes the experimental results obtained in estimating the error model and in applying it to error reduction. The significant error reductions observed in the automatic estimation of proportions in spoken surveys are particularly useful for a trend analysis of opinions over time.

2. Problem formulation

An opinion is defined as a pair $<x, \pi>$, where x represents the topic and $\pi \in \{\text{positive}, \text{negative}\}$ the polarity of the opinion. An opinion support is defined as a word string conveying an opinion $<x, \pi>$.

2.1. The opinion detection system

The architecture of the system used for detecting opinions is shown in figure 1:

- The segmentation process transcribes the oral message

[1]The work described in this paper was partially supported by the European LUNA project (FP6-33549) and by France Telecom grant N 021B178.

according to specific opinion LMs. Segments that are not detected with enough confidence are rejected. As a consequence, if no segment is kept for a message, it is rejected.

- The classification process associates to each reliable hypothesized opinion support the corresponding pairs $<x, \pi>$ with a probability score. A segment for which this association is not made with a sufficient confidence is interpreted as not being an opinion support. Furthermore, if there are no segments in a message labeled with an opinion hypothesis, then the message is considered as not expressing any opinion and will be taken into account only for the estimation of opinion proportions.

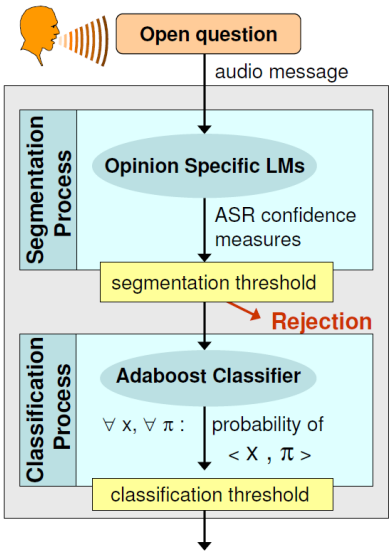

Figure 1: The opinion detection and interpretation system.

2.2. Opinion proportion analysis

There may be several opinion supports in one message expressing different polarities about the same topic. A global opinion value v is introduced to describe the global consolidated opinion $\{x, v\}$ expressed about the topic x in a message.

Let $Op(a, x)$ be a function returning a consolidated opinion value in a message a about a topic x. The consolidated opinion value v takes the following values : *none*, if the topic is not expressed; *satisfied*, if only positive opinion are express about the topic; *dissatisfied* if it is only negative polarities; and *mixed*} otherwise. The variable a takes the value m if the message has been manually segmented and each segment has been annotated with an opinion. The variable a takes the value $\widehat{m}$ if the message has been automatically segmented and annotated by the system whose architecture is represented in figure 1.

Let C be a corpus of n oral messages expressing opinions about a telephone service. Let $C' \subseteq C$ be a subset of C containing n' messages selected by the automatic interpretation strategy described in [3].

Let $p_C(x, v)$ be the reference proportion of messages m of the corpus C expressing the global opinion value v about the topic x. It is defined as:

$$p_C(x, v) = \frac{|m \in C \text{ and } Op(m, x) = v|}{n} \quad (1)$$

where $|g(m)|$ indicates the number of messages for which $g(m)$ is true.

Let $\widehat{p_{C'}}(x, v)$ be the proportion of automatically transcribed messages $\widehat{m}$ of the corpus C' expressing the global opinion value v about the topic x. It is defined as:

$$\widehat{p_{C'}}(x, v) = \frac{|\widehat{m} \in C' \text{ and } Op(\widehat{m}, x) = v|}{n'} \quad (2)$$

Different types of errors make $\widehat{p_{C'}}(x, v)$ different from $p_C(x, v)$. It is interesting to analyze the types of these errors to see if they can be modeled and partially corrected.

3. A model for survey errors

Rejecting a number of messages in a survey may lead to opinion proportions different from the ones computed with the full message population. The rejection of some messages corresponds to sampling the original message population with a sampling criterion based on confidence measures computed by the automatic ASR system as described in [3]. In order to separately characterize the errors due only to this sort of sampling and the ones due only to interpretation, it is useful to introduce $p_{C'}(x, v)$, the proportion of the consolidated message opinion $\{x, v\}$ obtained with the manual transcriptions and annotations of messages in the subset C' of C . It is computed as follows :

$$p_{C'}(x, v) = \frac{|m \in C' \text{ and } Op(m, x) = v|}{n'} \quad (3)$$

The difference between the true proportion $p_C(x, v)$ and the proportion $p_{C'}(x, v)$ is the first type of error defined as the sampling error :

$$err^{sampl.}(x, v) = p_C(x, v) - p_{C'}(x, v) \quad (4)$$

The other type of error in opinion proportion estimation is due only to interpretation errors and is defined as the interpretation error:

$$err^{interp.}(x, v) = p_{C'}(x, v) - \widehat{p_{C'}}(x, v) \quad (5)$$

The global error on the proportion estimation of opinion hypotheses is given by:

$$err^{global}(x, v) = err^{sampl.}(x, v) + err^{interp.}(x, v) \quad (6)$$

The interpretation error $err^{interp.}(x, v)$ depends on two types of errors, namely : *deletion* (d) occurring when a message m, contains $\{x, v\}$ which is not hypothesized in $\widehat{m}$ and *insertion* (i) occurring when a message m, does not contain $\{x, v\}$ which is hypothesized in $\widehat{m}$. Notice that substitution errors correspond to the deletion of the correct opinion and the insertion of another one at its place.

Let $d(x, v)$ be the deletion error rate and $i(x, v)$ be the insertion error rate for a consolidated opinion $\{x, v\}$. These errors depend on the opinion detection system and affect the estimated proportion with this quantity k(x,v) = d(x,v) - i(x,v). An error model is introduced by assuming k(x,v) to be constant. Based on this model, the following approximation of the error $\epsilon^{interp.}(x, v)$ due to the interpretation of the non-rejected messages is introduced:

$$\widehat{err^{interp.}}(x, v) = \widehat{p_{C'}}(x, v) * \frac{1 - k(x, v)}{k(x, v)} \quad (7)$$

Assuming also that the sampling error is constant and equal to $\widehat{err^{sampl.}}(x,v)$, the following approximation for the global error is considered:

$$\widehat{err^{global}}(x,v) = \widehat{err^{sampl.}}(x,v) + \widehat{p_{C'}}(x,v) * \frac{1-k(x,v)}{k(x,v)} \quad (8)$$

A prediction of the overall error $err^{global}(x,v)$ for opinion (x,v) is related in this way to the proportion $\widehat{p_{C'}}(x,v)$ estimated with the opinion hypotheses generated by the automatic system.

The estimations are not far from the reality as far as the performance of the opinion detection system is rather stable. Experiments described in the next section show that errors exhibit minor divergence with respect to a linear approximation of the equation 8. This suggests introducing a new estimation $\widehat{p_{corr,C'}}(x,v)$ of proportions as follows:

$$\widehat{p_{corr,C'}}(x,v) = \widehat{p_{C'}}(x,v) + \widehat{err^{global}}(x,v) \quad (9)$$

where $\widehat{err^{global}}(x,v)$ is obtained by the linear approximation of equation 8 for the observed proportion $\widehat{p_{C'}}(x,v)$.

The residual error, *i.e.* the reminded global error that persists after the correction, is obtained as follows :

$$err^{resid.}(x,v) = p_C(x,v) - \widehat{p_{corr,C'}}(x,v) \quad (10)$$

4. Experimental results

This section aims to compute an error correction function analyzing in details the different errors generated by the opinion detection system.

4.1. Data-gathering

France Telecom R&D collected several opinion survey corpora in order to analyze users satisfaction in a customer service. The results reported in this paper make use of one of them.

The corpus used, described in details in [3], contains 1354 calls. The corpus was manually transcribed and opinion supports were annotated with pairs $< x, \pi >$, with $x \in \{\text{efficiency}, \text{courtesy}, \text{rapidity}\}$.

The corpus has been split in two equal parts. The first corpus, denoted *TRAIN*, was used for estimating a linear approximation of the relation (8). The second corpus, denoted *TEST*, was used for evaluating the efficiency of error correction. These data were exclusively used for the test.

In order to have appropriate data for estimating the parameters of the linear approximation, the train and test corpora were manually split into a set of specific sub-corpora. They were constructed for each opinion by ramdomly picking samples to have different proportions in a sufficiently large interval. In order to accomplish this task, the number and size of resulting sub-corpora are specific to each opinion.

For the sake of brevity, detailed results will be given for two polarities of the most frequently expressed and most interesting opinion topic, namely *efficiency*. The distributions of $p_C(\textit{efficiency}, \textit{dissatisfied})$ for each corpora are from 0.23 to 0.42 and from 0.33 to 0.51 for $p_C(\textit{efficiency}, \textit{satisfied})$.

4.2. Error estimation

For each set in the *TRAIN* and *TEST* corpora, the proportions $p_C(\textit{efficiency}, \textit{satisfied})$ and $p_C(\textit{efficiency}, \textit{dissatisfied})$ were computed. The proportions $\widehat{p_{C'}}(\textit{efficiency}, \textit{satisfied})$ and $\widehat{p_{C'}}(\textit{efficiency}, \textit{dissatisfied})$ were estimated from spoken messages with the opinion detection system. The observed errors for each set was computed as defined in equation 6.

The overall error $err^{global}(x,v)$ for each of the six sets in the *TRAIN* corpus and their linear interpolation function of these data points are plotted as function of $\widehat{p_{C'}}(x,v)$ in figure 2 for $\{\textit{efficiency}, \textit{satisfied}\}$, while figure 3 shows the same type of data for $\{\textit{efficiency}, \textit{dissatisfied}\}$.

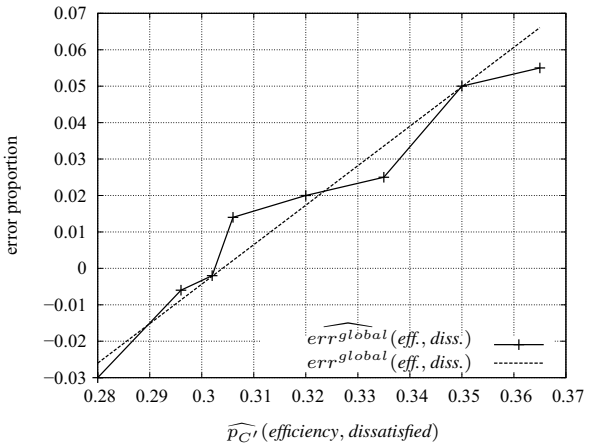

Figure 2: Estimation and exact value of the overall error according to the hypothesized proportions $\widehat{p_{C'}}(x,v)$ for x =*efficiency* and v =*dissatisfied* on the corpora randomly sampled from *TRAIN*

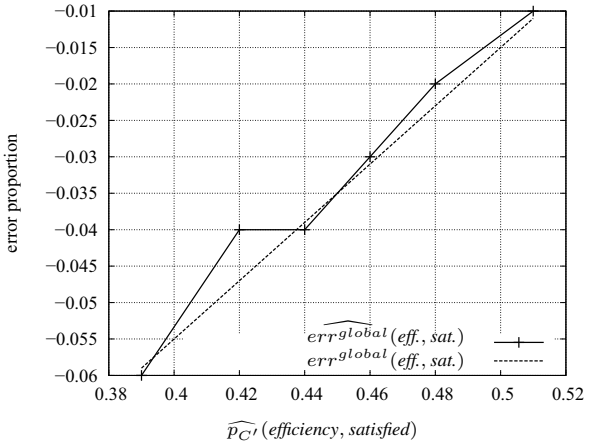

Figure 3: Estimation and exact value of the overall error according to the hypothesized proportions $\widehat{p_{C'}}(x,v)$ for x =*efficiency* and v =*satisfied* on the corpora randomly sampled from *TRAIN*

4.3. Error correction

The linear interpolation function estimated with the data in each figure has been used as a model for error correction. For each set of the *TEST* corpus, the proportion $\widehat{p_{C'}}(x,v)$ has been used for retrieving a value of the linear function. This value has been used for estimating $\widehat{p_{corr,C'}}(x,v)$. The results of error correction are shown in figures 5 for $(\textit{efficiency}, \textit{satisfied})$ and 4 for $(\textit{efficiency}, \textit{dissatisfied})$.

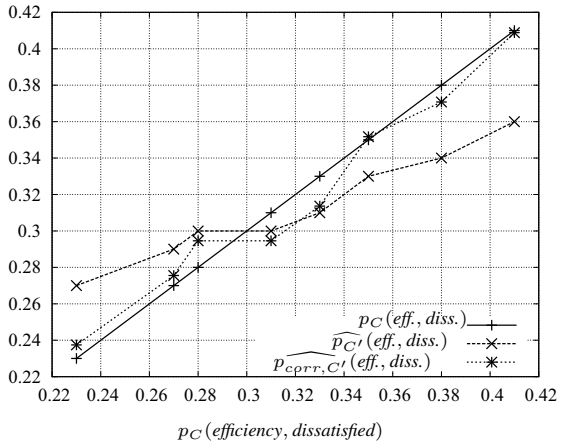

Figure 4: Evaluation of the error proportion correction applied to (*efficiency*, *dissatisfied*)

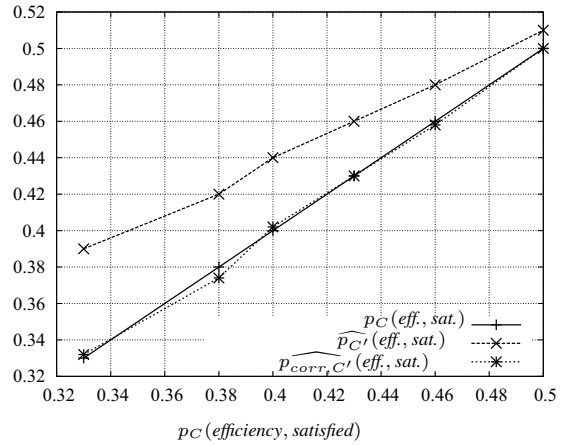

Figure 5: Evaluation of the error proportion correction applied to (*efficiency*, *satisfied*)

Table 1 reports, for the TEST corpus, the variances and confusion intervals for the global error and the residual error computed after applying bias correction.

x =*efficiency* and v =	*satisfied*	*dissatisfied*
variance($err^{global}(x,v)$)	0.0001	0.0009
confidence interval	0.0175	0.0578
average ($err^{global}(x,v)$)	-0.03	0.075
maximum ($err^{global}(x,v)$)	-0.05	0.05
After bias error correction		
variance($err^{resid.}(x,v)$)	0.0000	0.0001
confidence interval	0.0125	0.0198
average ($err^{resid.}(x,v)$)	-0.0014	-0.017
maximum ($err^{resid.}(x,v)$)	-0.0107	0.0132

Table 1: Estimated values for the *efficiency* topic with values v =*satisfied* and v =*dissatisfied* obtained on *TEST*.

5. Conclusion

The types of errors made by a system for performing automatic surveys of spoken opinions have been analyzed. The analysis of these errors in the estimation of opinion proportions has shown that there exists a relation between the estimated proportions and the true ones. A linear approximation between the estimation error and the estimated opinion proportions can be obtained with a development corpus. A model for the proportion estimation errors based on the linear approximation has been derived and used for partially correcting the errors on the automatic estimation of proportions. Experimental results on a test set have shown a significant reduction in the estimation errors and their variances making this approach suitable for monitoring proportion variations in time and with them the quality of service. After computing the confidence intervals, it is possible to conclude that a variation of 0.02 after error correction (as opposed to 0.06 before correction) in the estimated proportion of dissatisfaction is sufficient for assessing with confidence a decrease in satisfaction.

6. References

[1] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

[2] Nathalie Camelin, Frederic Bechet, Geraldine Damnati, and Renato De Mori. Automatic customer feedback processing: alarm detection in open question spoken messages. In *Interspeech*, Brisbane, Australia, Septembre 2008.

[3] Nathalie Camelin, Geraldine Damnati, Frederic Bechet, and Renato De Mori. Detection and interpretation of opinion expressions in spoken surveys. *IEEE Transactions on Speech and Audio Processing*, accepted with minor revisions, 2009.

[4] W.G Cochran. *Sampling Techniques*. John Wiley and Sons, 1977. Third Edition.

[5] J. Grothendieck and A. Gorin. Towards link characterization from content. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 4849–4852, 2008.

[6] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics.

[7] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *HLT/EMNLP*, 2005.

[8] Jayce Wiebe, Theresa Wilson, and Claire Cardie. Annotationg expressions of opinions and emotions in language. In *Language Resources and Evaluation (formely Computers and the Humanities)*, volume 39, pages 165–210, 2005.

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