INTERSPEECH 2009 BRIGHTON

Automatic Topic Detection of Recorded Voice Messages

Caroline Clemens[1], *Stefan Feldes*[2], *Karlheinz Schuhmacher*[1], *Joachim Stegmann*[1]

[1] Deutsche Telekom Laboratories, Berlin, Germany, [2] T-Systems, Darmstadt, Germany

caroline.clemens@telekom.de, stefan.feldes@t-systems.com,
karlheinz.schuhmacher@telekom.de, joachim.stegmann@telekom.de

Abstract

We present an approach to automatic classification of spontaneously spoken voice messages. During overload periods at call-centers customers are offered a call-back at a later time. A speech dialog asks them to describe their concern on a voice box. The identified topics correspond to the supported service categories, which in turn determine the agent group the customer message is routed to. Our multistage classification process includes speech-to-text, stemming, keyword spotting, and categorization. Classifier training and evaluation have been performed with real-life data. Results show promising performance. The pilot will be launched in a field test.

Index Terms: speech analytics, speech-to-text, automatic call processing, call center

10.21437/Interspeech.2009-264

1. Introduction

Research and development of automatic classification in speech dialog systems has been pushed forward in recent years. A lot of this work aims to classify by voice or personal features like age or gender [1]. Automatic topic detection, however, requires classification with respect to the spoken content. The classification presented in this paper is part of a framework that aims at enhancing customer service. The need for our topic detection approach arises from an analysis of shortcomings in telephone-based customer service. It is common practice in call centers that, in case of overload, a calling customer is left on hold and presented with music or commercial announcements. This situation is often unsatisfactory. Interactive voice response systems have been introduced with the aim of handling simple standard situations that do not require the involvement of service personnel, e.g. in telephone banking. Other approaches to reduce call overhead time for both customer and call center agents include automatic call routing. The goal is to directly route the customer to the appropriate service personnel based on the analysis of the user's fluently spoken response to some open-ended system prompt like ''How may I help you?'', e.g. [2]. If, however, all respective service personnel is busy the customer will still remain waiting.

In our approach the customer is offered to leave a message on a voice box describing his/her concern. The customer will then be called back in lower traffic periods when a service agent is available. In order to allow for a qualified call back the recorded message has to be classified into one or more of the possible service categories and routed to the corresponding skill group of call center agents.

A similar classification task, however based on textual input, has already been addressed in [3]. With the application background of an online community the goal of that classification approach was to find a suitable expert for a given problem based on a free text user request.

Classification in a call center environment has additionally to cope with various effects of spontaneous speech, [4] including emotional aspects. Furthermore, the concern described by the customer can often not be assigned unambiguously to the existing service categories. This may not only happen in the case of vague wording caused by the spontaneity of the user answer. Clearly formulated statements may also lead to more than one category. At the same time, a misclassification may be uncritical as long as the customer's concern can be handled by the addressed skill group. This issue must be reflected when developing the categorization scheme (s. section 3). Such more or less critical misclassifications should be handled appropriately in the evaluation of the classification. In section 4 we therefore discuss a suitable evaluation metric.

The paper is organized as follows. After a description of the application scenario we derive the requirements for the classification. We then describe the category development and the classifier integration into the overall system. The proposed classification is evaluated with real-world data and results are presented. We conclude with a discussion and outlook.

2. Application scenario

The scenario flow is visualized in Figure 1.

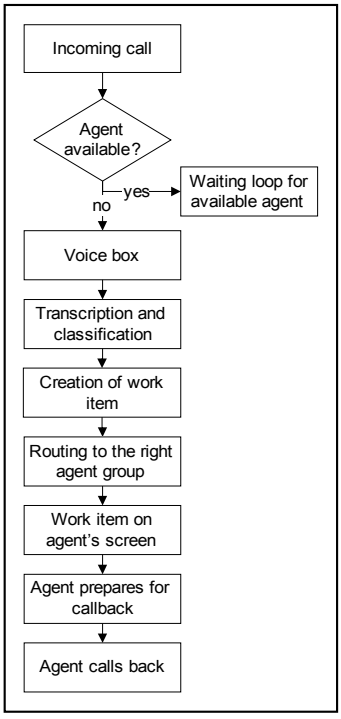

Figure 1. *Scenario flow from incoming call to call-back.*

We consider the situation when all call center agents of a customer service are busy. As an alternative to waiting, a calling customer is offered the option to leave a message with her/his concern on a voice box for a later call-back. The recorded message is classified according to the supported service categories and an agent of the corresponding skill group can handle the work item once he or she is available again.

To provide information to the agent in a clearly represented form the user utterance is automatically transcribed by a speech-to-text unit and displayed to the agent together with highlighted category-relevant text passages found by the classifier. In addition, there is still the option to listen to the recorded customer message. The information offered allows the agent to understand the customer's concern at a glance and to open relevant documents and specific customer files while setting up the call-back.

Beyond this primary use case the transcription and classification result may also be used for process quality management. It could, for example, be checked whether the number of messages classified to the "billing" category increases after a change in the billing system. Use cases could also be realized for marketing observation – for example automatically counting how often a new product name was recognized after a specific marketing campaign.

3. Speech Analytics

3.1. Overview

The process flow of the multistage classification from the incoming voice recording to the classifier output is shown in Figure 2. An audio recording of the customer messages is transcribed by a speech-to-text module, processed by a stemming algorithm, and classified by the automatic classifier which creates a text file containing the results as output.

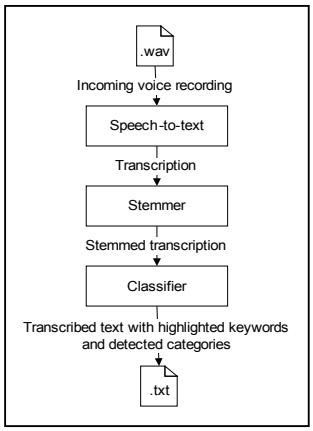

Figure 2: *Process flow from voice recording to output text of the classifier.*

3.2. Speech Corpus

To develop and train the speech analytics functionality, including speech-to-text and automatic classification, it was necessary to adequately prepare the speech data. Our corpus consists of more than 5000 original customer recordings. These recordings derive from customers who called a service hotline and left a voice message on a mailbox. The recording time was limited to 60 seconds. Only a few customers made use of the complete recording time, the mean length of the recordings is around 12 seconds. The content varies from single words to several sentences, partly including emotional utterances.

3.3. Speech-To-Text Tuning

The incoming speech material poses challenges for speech recognition, as it contains a large number of heterogeneous speakers with various dialects and accents. Customers speak as spontaneously as they are used to when leaving a message on an answering machine. The voice recordings naturally exhibit many effects associated with spontaneous speech. Some customers speak very quickly or emotionally (anger). In order to optimize the speech-to-text system to these conditions, manually-transcribed training material was used. Detailed transcription rules for the human transcribers were developed to include marks for unconventional pronunciations or noise events. The resulting hand-transcribed data with its accuracy and reliability formed a rich input for the training of a domain-specific lexicon and domain-specific acoustic model. After this optimization, the speech-to-text system attained a word error rate of about 25%.

3.4. Taxonomy development

In our recordings customers report a concern. A viable taxonomy must provide categories that the customer concern can be sorted to. For the first version of the taxonomy we chose a top-down approach. It consists of categories that reflect the structure of processes established in the call center workflow. The recordings were manually sorted by the reported customer concern. This first version of categories did not turn out to be a practicable taxonomy, as the distribution of recordings to categories was quite unbalanced. Furthermore, some recordings were difficult to sort, as none of the categories fitted the customer concern.

To improve the taxonomy we combined the top-down approach with a bottom-up approach. The top-down approach contains established support categories that are mapped to respective skills of the call center agents. These categories should be clearly distinguishable by the wording that customers use when they describe their concern. Therefore, the bottom-up part of the new approach was based on an analysis of the collected recordings. We sorted the reported questions of the costumers by their content and then grouped related concerns. Combining both approaches we achieved a taxonomy that considered both the reported concerns and the existing support categories. The transcribed recordings were sorted to the categories by hand. The combined approach leads to a taxonomy that showed a more consistent distribution. Table 1 shows the new taxonomy.

Category		Subcategory	
A	Sales		
B	Customer Care	c	Engineer
		d	Device
		e	Awaiting
		f	Move
G	Network or Connection Breakdown		
H	Billing		
M	Miscellaneous		

Table 1. *Taxonomy of the categories (Text in this table has been translated into English for comprehensibility).*

The categories represent the following topics: Category A "Sales" covers all concerns about contracts; Category B "Customer Care" is further characterized by a subcategory layer. However, category B is not only the sum of its subcategories, but moreover covers those concerns that may not be assigned to any of its subcategories; "Engineer" contains topics related to an appointment with a technician at the customers' home; "Device" stands for all questions about terminal equipment; "Awaiting" refers to concerns about ordered services the customer is waiting for; "Move" covers topics about moving to another address; Category G "Network or connection breakdown" includes questions and reports on network problems; H "Billing" stands for all billing matters. Category M "Miscellaneous" comprises any other message that can not be assigned to one of the categories above.

3.5. Automatic classifier

The classifier can handle taxonomies with several layers [3], as the taxonomy architecture is a tree in which a node can contain several subnodes. This architecture allows easy modification of the taxonomy if necessary. The results of the classification process are one or more suitable categories for each transcribed recording.

In our scenario (s. section 2) output texts of the speech-to-text system are the input for the automatic classifier. These texts are first processed by a stemming algorithm to remove typical German prefixes and suffixes (Figure 3). The used stemming process is strictly rule-based and cuts off predefined letters at the beginning and end of a word.

was passiert wenn ich einen DSL Anschluss Call und Surf Komfort vor Ablauf der Windes Antrags Laufzeit kündige ich glaube irgendwo mal gelesen zu haben da muss man eine Ablöse bezahlen in ungefähr 5 und 20 Prozent vom Gesamtpreis

Figure 3. *Example of speech-to-text input for the classifier. Grey highlighted affixes are deleted by the stemmer (Text in this figure was left in German to point the affixes occurring in German language).*

As training material for the automatic classifier, we used transcriptions of the speech-to-text engine that were assigned to the proper categories by hand. We chose this method to provide the classifier with the same quality of domain-specific data as the later original input. Each recording could be labeled to more than one category, as some recordings contained combinations of topics.

In the classifier training phase a statistical analysis computes words in the training material that are typical for the given categories. For details of the classifying process see [5]. To improve the classifier performance experts of the domain defined an additional set of keywords for each category that have higher priority in the classification procedure: if one of these keywords is found, the recording is sorted to the relevant category. This so-called hybrid approach combines the statistically-derived keywords with the expert-defined keywords.

The classifier output consists of three parts (Figure 4): the transcribed text of the voice message, highlighted relevant keywords in the transcribed text, and the detected topics. This output is used to create a so-called work item (a described working task) for a call center agent. Further information that has been gathered in a preceding IVR dialogue, e.g. customer phone number, customer id, call time, can be added.

I'm moving house next month and I want you to change my address so the bills are sent to the new address from now on

Categories: MOVE, BILLING

Figure 4. *Example of the output of the automatic classifier (Text in this figure has been translated to English for comprehensibility).*

4. Evaluation and Tuning

4.1. Setup

For a first assessment, the necessary manual preparation of training material for the tree-based statistical classifier was only performed for a subset of customer messages from the corpus described in section 3.2. A training corpus of 573 messages was transcribed and labeled. Accordingly, a test set of 242 messages was formed.

4.2. Metrics

For text classification as well as for other text comparison tasks, the recall, precision, and F-measures are well-established and widely use [6]. However, they do not naturally reflect that different errors may have different impact on the application. In our case, it is essential that the addressed call center agent has the required skills to handle the respective customer concern. The quality figure calculated by some suitable metric should therefore correlate with the quality perceived by the human agent. Thus, our first objective was to find a metric that reflects the pragmatic differences between categories appropriately in our context. We first experimented with a self-developed, heuristic, table-based distance measure. It worked satisfactorily, but as it is task dependent it requires manual rebuilding if changes in the taxonomy occur. This turned out impractical during the development phase.

Due to the fact that we could organize our taxonomy in a hierarchical form, it was possible to set up a fully computational method in quite a straightforward way. We applied an expansion strategy to the hypothesis and reference items (classifier output and expected result) before calculating the standard recall and precision metric. This method penalizes misclassifications between subcategories less than misclassifications between main categories. The implemented expansion rules are: For any child category in a reference item, add the parent category, if this parent category or a corresponding sibling category is part of the hypothesis item. Accordingly, for any child category in a hypothesis item, add the parent category, if this parent category or a corresponding sibling category is part of the reference item. Table 2 illustrates the expansion strategy, where X, Y are two main categories and a, b are two child categories of parent X whereas c is a child category of Y.

Original				Expanded			
Ref	**Hyp**	**Rec**	**Pre**	**Ref**	**Hyp**	**Rec**	**Pre**
a	X	0.0	0.0	X, a	X	0.5	1.0
a	b	0.0	0.0	X, a	X, b	0.5	0.5
X	a	0.0	0.0	X	X, a	1.0	0.5
a	a	1.0	1.0	a	a	1.0	1.0
X	Y	0.0	0.0	X	Y	0.0	0.0
a	c	0.0	0.0	a	c	0.0	0.0

Ref: Reference item
Hyp: Hypothesis item
Rec: Recall value
Pre: Precision value

Table 2. *Illustration* of the expansion strategy.

4.3. Results

The results shown below were obtained using the tree-based statistical classifier. First, a test was performed on the reference set to check how the classifier has adapted to the training set. Values of 99 % for recall and 88% for precision indicate satisfactory performance. The results of the independent test set are shown in Table 3. Regarding the unequal distribution of category occurrences, we first calculate the recall and precision values for each category separately and obtain the total values by calculating the respective frequency weighted sum. For some categories we obtain reasonable performance. However, the majority of categories suffers from sparse references in the test as well as in training data and therefore tends to have low recall and precision values. Obviously, results suggest improvement by enlarging the training set. In addition, it must be taken into account that the data used in this evaluation was derived from an application context that was not completely identical to the scenario we envisage in this paper. We therefore set up a prototype test (s. section 4.4).

Category	Freq. of Occurrence	Recall	Precision
M	132	0.86	0.59
A	58	0.55	0.52
B	14	0.71	0.91
c	7	0.14	0.33
d	8	0.25	0.50
e	28	0.14	0.33
f	2	0.50	1.00
G	20	0.65	0.68
H	16	0.75	0.67
Total	285	0.66	0.57

Table 3. *Evaluation result for statistical classifier.*

Preliminary tests have also been performed on the hybrid classifier approach that additionally includes an application specific keyword lexicon as described in section 3.4. First results are promising and indicate improvements in domain specific scenarios.

4.4. Prototype Test

In order to get more specific data for tuning the classifier we are currently running a friendly-user-test: Users are asked to choose a customer service scenario, call the prototype system and leave a voice message describing the chosen concern in their own words. In response they immediately receive an email with a link to a web page on which the result of the classification is presented together with the transcription of the message produced by the speech-to-text engine. The transcription is annotated by highlighting those text passages that lead to the classification result (compare Figure 4). Unlike the real application, the user in this test is asked to give feedback by confirming or correcting the classification result. Additionally, the user can rate the quality of the automatic transcription. As a result of this test-run we expect a valuable enlargement of the training corpus allowing for substantial classifier tuning. Furthermore, it will provide subjective performance ratings as a supplement to the metric evaluation results.

5. Conclusions

The primary aim of the presented topic detection in a customer service environment is to route a recorded customer message to a human agent with the specialized skill to solve the customer concern. The proposed concept of a classifier based on a hierarchical taxonomy proved its principal suitability for this purpose. We developed such a taxonomy, containing the customer concerns of the given domain. For a first training and evaluation, real data recorded from a customer service hotline was prepared. However, for further tuning more training data specific to the envisaged scenario is needed. Therefore, based on the prototype system a friendly user test has been set up and a pilot will be launched in a field test in 2009.

6. Acknowledgements

The authors would like to thank the project team and the members of our project partners for their help, namely our partner project Spree and the cooperating DAI-Lab.

7. References

[1] Burkhardt, F., Metze, F., and Stegmann, J., “Speaker Classification for Next Generation Voice Dialog Systems.” in R. Martin, U. Heute, C. Antweiler. [Ed], Advances in Digital Speech Transmission, Wiley, 2007.

[2] Gorin, A., Riccardi, G., and J. Wright, “How may I help you?”, Speech Communication, vol. 23, 113–127, 1997.

[3] Metze, F., Bauckhage, C., Alpcan, T., Dobbrott, K., and Clemens, C., “The spree expert finding system”, in Proc. of ICSC, First IEEE International Conference on Semantic Computing, Irvine, 2007.

[4] Camelin, N., Béchet, F., Damnati, G., and Mori, R. De, "Speech mining in noisy audio message corpus", INTERSPEECH-2007, 2401-2404, 2007.

[5] Wetzker, R., Alpcan, T., Bauckhage, C., Umbrath, W., and Albayrak, S., “An unsupervised hierarchical approach to document categorization”, IEEE Intl. Conf. on Web Intelligence (WI’07), Silicon Valley, USA, 2007.

[6] Yang, Y., and Liu, X., “A Re-examination of Text Categorization Methods”, in Proc. SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, 1999.