Automatic Categorization of Voicemail Transcripts Using Stochastic Language Models

Konstantinos Koumpis

Vienna Telecommunications Research Center - ftw.
Tech Gate Vienna, 1 Donau City St., Vienna 1210, Austria
Email: koumpis@ftw.at

Abstract. This paper is about the applicability of stochastic language models to the task of categorizing voicemail message transcripts. The target categories are related to priority and content and are thus suitable for mobile messaging applications based on profiles which can be determined by users' physical and social environment. Categorization is performed by comparing the posterior probabilities of test messages under the language models of each target category. Stochastic models were selected over other lexical features because of their ability to incorporate context dependencies while their parameters are determined automatically from data. Despite the relatively small amount of training data used and given the spontaneous nature of voicemail, the models performed fairly accurately. Our experiments examine the effects that factors such as the word error rate, the n-gram order, smoothing and textual representation have on overall categorization accuracy.

1 Introduction

Voicemail represents a significant amount of spoken audio stored daily in digital form as a byproduct of telecommunications systems. Voicemail features a conversational interaction between a human and a machine with no feedback from the machine and for which manual organization is a time consuming task, particularly for high-volume users. There are situations in which users would prefer to receive messages of certain content types and leave the remaining ones to be reviewed later at a more convenient location or time. Today, voicemail recipients rely almost exclusively on caller line identity – the display of caller's phone number or name – to screen incoming messages.

A few alternative solutions have been proposed for efficient voicemail retrieval and management which include browsing and searching of message transcriptions via a graphical user interface [1], generation of text summaries for wireless handheld devices [2], extraction of caller identity and phone number from messages [3], and message ranking based on urgency and business relevance [4]. A message categorization system can instead sift through a stream of arriving messages to find those relevant to a user profile. Unlike search queries, user profiles are persistent, yet adaptive, and tend to reflect a long term information need. Considering a general *voicemail categorization* task, each spoken message can be assigned to none, one or multiple predefined categories.

Constructing and maintaining rules with reasonable complexity for categorization is a tedious and possibly not robust task, if unrestricted domains and spontaneous speech input are to be targeted. It is possible instead to build classifiers automatically by learning the

Petr Sojka, Ivan Kopeček, and Karel Pala (Eds.): TSD 2004, LNAI 3206, pp. 363–370, 2004.

characteristics of the categories from a training set of pre-classified examples. Many standard machine learning techniques have been applied to automated text categorization problems, such as decision trees, naive Bayes, neural networks, k-nearest neighbour classifiers and support vector machines [5,6,7]. The above approaches are effective when the texts to be categorized contain sufficient numbers of category specific terms so that a 'bag-of-words' model, which is based on a histogram of weighted word frequencies, can discriminate among the categories. Our pilot experiments with the Rainbow text categorization system [8] indicated, however, that probabilistic classifiers with isolated words as input features are not sufficient to perform voicemail categorization.

Stochastic language models are more robust than isolated words because they incorporate local dependencies as a result of modelling symbol sequences within the framework of standard Markov based approximations. Character level language models have been found to be effective in text classification [9] and author attribution [10] tasks. The present paper deals with a relatively small corpus of spoken message transcripts. These are different from written language as they are often ungrammatical, lack punctuation and capitalization, and almost always contain substitution, deletion and insertion errors. The rest of the paper is divided into five sections. Sections 2 and 3 describe, respectively, the voicemail corpus and the categorization protocol used. Section 4 discusses the methodology employed to perform message categorization, and section 5 gives the experimental results. Finally, we summarize our conclusions and discuss future work in Section 6.

2 Voicemail Corpus

We have used the LDC Voicemail Corpus-Part I [11]. This corpus contains 1801 messages (14.6 hours, averaging about 90 words per message). As a training set for the categorization tasks we used 1000 messages (84K words) from this corpus (messages 1 to 800 and 1602 to 1801). For evaluation purposes we used the test set of the corpus comprising 42 messages (2K words) as well as the test set of the Voicemail Corpus-Part II comprising 50 messages (4K words). Apart from the human transcriptions (denoted **SR-HT**), which contained some noise in the form of repetitions and broken words, we also used transcriptions with a word error rate (WER) of 42.5% produced by a hybrid multi-layer perceptron / hidden Markov model speech recognizer (denoted **SR-SPRACH**) [12]. Additionally, we obtained another set of transcriptions with a WER of 31% (denoted **SR-HTK**) produced by the more complex HTK Switchboard system adapted to voicemail [13].

3 Voicemail Categorization Protocol

For the automatic categorization of voicemail messages we consider two tasks, categorization by *priority* and by *content*. The categories in both tasks are mutually exclusive and exhaustive, that is, every message belongs to one, and only one, of the categories. The data labelling is a result of subjective analysis of the message transcriptions. The attributes that a message recipient will perceive along with the categorization criteria, are determined by individual needs.[1] These needs change over time and with the physical and social environment. As the

[1] Theories exist in order to understand the ways humans categorize objects [14].

Table 1. Taxonomy for the message priority- and content-based categorization tasks.

Priority-based categorization

Category	Description	Examples
high	an immediate action by the recipient is required, expected or implied (often following a request)	visitor waiting, meetings rescheduling, provide help or components
medium	some attention by the recipient will be required	parcel deliveries, lunch invitations, project updates
low	rather trivial content, no need for immediate attention	greetings, jokes, shopping, tickets for sport/leisure events

Content-based categorization

Category	Description	Examples
technical	specific technical issues related to projects	technical problems or solutions related to ongoing projects, software programming
office	daily issues (excl. technical)	equipment maintenance or upgrade, arrival of faxes and parcels (excl. personal) printing tasks, help-desk or security responses
business	complementary professional tasks not covered by the above	meeting schedules, reminders and availability, customer visits/demos, recruitment, corporate marketing, purchase orders, travel arrangements (excl. personal)
family	related to family members (spouse, children, parents etc.) or concern family issues	availability, shopping list, homework, guests
friends	related to friends (incl. colleagues but not concerning work)	sports and leisure activities, availability
private	miscellaneous content concerning the recipients not covered by any of the above	bookings of restaurants or holidays, medical appointments, delivery of items (excl. work)

corpus is not organized per voicemail subscriber, we assumed a general voicemail recipient profile, which might not be fully compatible with the criteria of each individual voicemail recipient. During the labelling process for the categorization tasks no attempt was made to associate the message priority or content with the identity of speakers and thus the task does not share similarities with speaker recognition [15].

Table 1 outlines our taxonomy along with examples related to the priority- and content-based categorization tasks. Given the relatively small size and the nature of the corpus, we decided to use 3 and 6 categories, respectively, because in a dense category space there would be only a few example messages in each category. The distribution of messages in the training and test sets for the priority- and content-based tasks are given in Fig. 1.

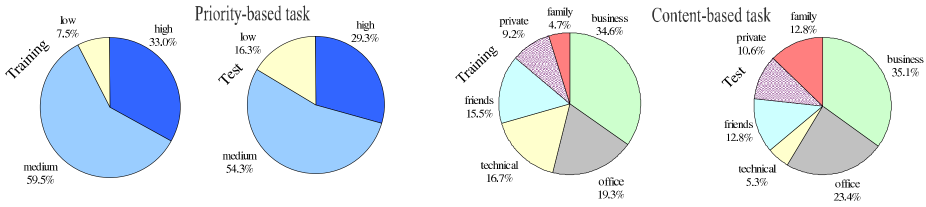

Fig. 1. Category distributions across the training and test sets related to the priority (left) and content (right) tasks, respectively.

4 Categorization Using Stochastic Language Models

Stochastic language models attempt to capture regularities of natural language for the purpose of improving the performance of various language engineering tasks [16]. Probabilities are assigned to linguistic symbols (e.g., words, syllables or characters) and mathematical models are used to represent statistical knowledge. The probability of a sequence $s_1, s_2, \ldots, s_i$ is given by:

$$p(s_1, s_2, \ldots, s_N) = \prod_{i=1,\ldots,N} p(s_i | s_1, \ldots, s_{i-1}) \tag{1}$$

A simple yet effective approach to approximate the above is provided by n-gram models according to which the occurrence probability of any test symbol sequence is conditioned upon the prior occurrence of $n - 1$ other symbols:

$$p(s_i | s_1, \ldots, s_{i-1}) \approx p(s_i | s_{i-n+1}, \ldots, s_{i-1}) \tag{2}$$

n-gram language models have the advantage of being able to cover a much larger variation than would normally be derived directly from a corpus in the form of explicit linguistic rules, such as a formal grammar. Open vocabularies can also be easily supported by n-gram language models.[2] Stochastic language models are usually employed in the context of Bayesian decision theory. The task of classifying a message transcription $\mathcal{M}$ into a category $c \in C = \{c_1, c_2, \ldots, c_C\}$ can be expressed as the selection of the category which has the largest posterior probability given the message transcription:

$$\begin{aligned} c^+ &= \arg\max_{c \in C} \{p(c|\mathcal{M})\} && (3) \\ &= \arg\max_{c \in C} \{p(\mathcal{M}|c)\, p(c)\} && (4) \end{aligned}$$

In the above expression the language model is used to estimate the likelihood $p(\mathcal{M}|c)$, whilst the prior $p(c)$ is assumed to be the same with that of the training set. For computational reasons, the product of probabilities in Eq. (4) is replaced by a sum of negative log

[2] For instance, we obtained consistently better results by mapping all out-of-vocabulary words to a single symbol. Thus all experiments reported in section 5 made use of open vocabularies.

probabilities. Categorizing a message involves calculating a sum of negative logs for each category, where the length of each sum equals the number of n-grams contained in the test message. The most likely category c^+ is then the one minimizing that sum. If one assumes equal priors this becomes equivalent to the perplexity criterion [17]. Comparing the above measure across different categories for each test message allows the highest ranked category along with a rank value to be returned.

5 Experimental Results

Categorization performance in all subsequent experiments is measured in terms of *overall accuracy*, which we define as:

$$Acc = \frac{\#correctly\ categorized\ messages}{\#messages\ considered} \qquad (5)$$

We examined the effects of the following factors on the above metric:

WER quantifies the mismatches between the reference category language models and those of the test messages due to transcription errors.

n**-gram order** introduces a trade-off between capturing enough context and having poor model estimates due to data sparsity. All models back-off to lower order n-grams.

smoothing replaces the original counts with modified counts so as to redistribute the probability mass from the more commonly observed events to the less frequent and unseen events. Various smoothing techniques were compared, namely linear, absolute, Good Turing and Witten Bell [18].

symbol types compare transcriptions which contain word strings (denoted **plain**), word strings subject to linguistic stemming (denoted **stem**) [19], word strings after removing 56 frequently occurring and semantically light functional words (denoted **stop**), and separate characters including spaces between words (denoted **char**).

The results for the priority- and content-based tasks are given in Figs. 2 and 3, respectively. Note that the training set is the same, whether we test on manually or automatically transcribed data.[3] The priority-based categorization task proved to be easier to perform than the content-based, although this may be due to the different degree of confusability (3 vs. 6 target categories) between the two tasks. As was expected, transcription errors had a significant impact on categorization accuracy. Moving from manual to automatic transcriptions with WERs of either 31% or 42.5% reduces the accuracy by about 20% absolute, across both categorization tasks. Plain textual representation offered higher accuracy than stemming, but the differences are smaller the higher the WER is. Removing stop words from the transcriptions led to consistently lower categorization accuracy, suggesting that frequently occurring and semantically light functional words play an important role in capturing differences among categories. Character-based n-grams were not as robust as word-based n-grams except for high WER conditions. Optimal n-gram order depended on the type of textual representation. In word-based representations trigrams

[3] We expect that the performance when testing with automatically transcribed data can be improved by using training data that is automatically generated too.

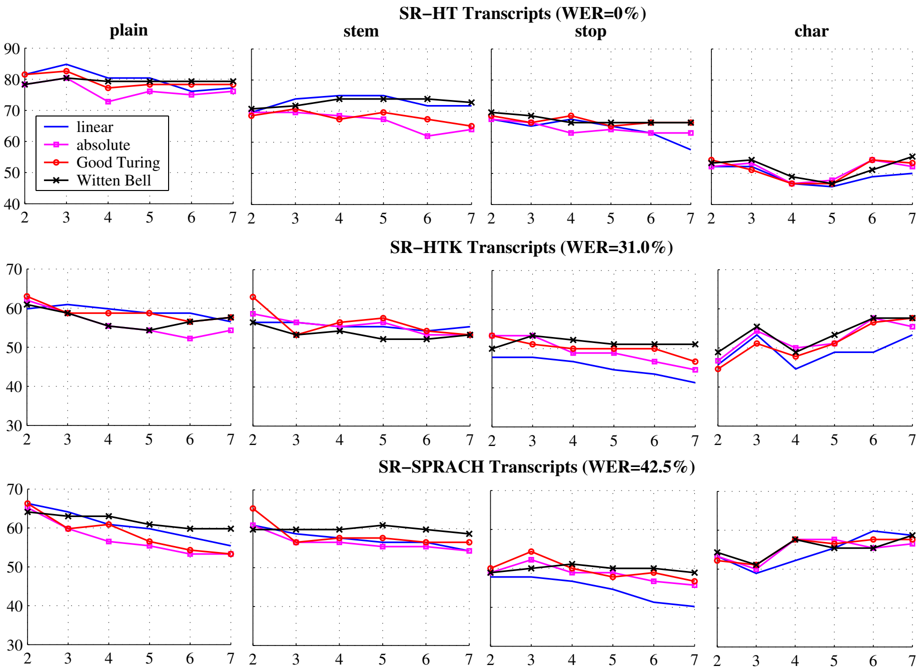

Fig. 2. Accuracy (%) in the priority-based categorization task using different smoothing techniques. The rows of subfigures correspond to transcripts of different WERs, while the columns correspond to different textual representations. The n-gram order is shown on the horizontal axis.

gave the best results when tested with manual transcriptions, while either bigrams or trigrams were optimal when automatic transcriptions were considered. In character-based representation optimal n-gram values were found to be in the upper range ($n = 6, 7$). Finally, small differences were observed in the results from each of the four smoothing techniques evaluated. Among them, linear and Witten Bell performed slightly better on average, followed by Good Turing. Linear smoothing was, however, occassionaly less robust to an increase in the n-gram order.

6 Concluding Remarks

The ability to categorize spoken messages into predefined categories using supervised learning has important applications in information retrieval, information filtering, and knowledge management systems. This paper has defined the task of voicemail categorization and presented a series of experimental results based on comparisons of stochastic language models. According to this approach training is performed by updating n-gram counts and categorization by comparing the normalized sum of the n-gram counts corresponding to the symbols in each test message adjusted by the prior of each category. Hence, training and categorization are both simple and efficient and can be easily integrated into a profile assisted

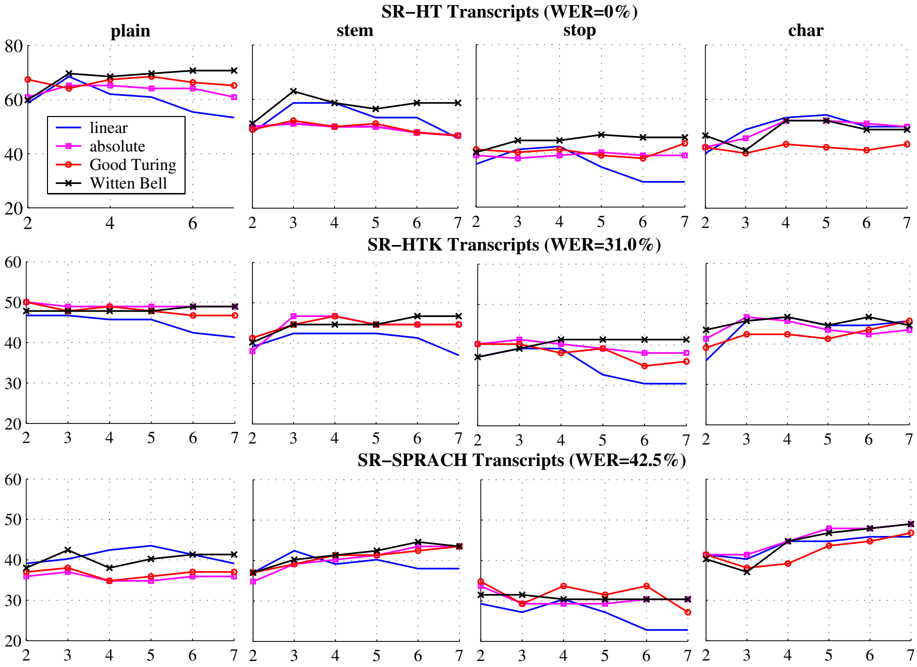

Fig. 3. Accuracy (%) in the content-based categorization task. The subfigure layout follows that of Fig. 2.

voicemail management tool. Current work involves supervised and semi-supervised training using larger sets of voicemail transcripts and the development of a methodology according to which the coverage of language models used to categorize messages can be augmented with information derived from statistics estimated from multiple corpora. An investigation into the applicability of the maximum entropy framework to the voicemail categorization task is also under way.

Acknowledgements

Georg Niklfeld and Ed Schofield provided valuable comments on earlier drafts of this paper. This research is supported by a Marie Curie fellowship.

References

1. Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Whittaker, S., Zamchick, G.: SCANMail: Browsing and searching speech data by content. In: Proc. EuroSpeech, Aalborg, Denmark (2001).
2. Koumpis, K., Ladas, C., Renals, S.: An advanced integrated architecture for wireless voicemail retrieval. In: Proc. 15th IEEE Intl. Conf. on Information Networking, Beppu, Japan (2001) 403–410.
3. Huang, J., Zweig, G., Padmanabhan, M.: Information extraction from voicemail. In: 39th Annual Meeting of Assoc. for Comp. Linguistics, Toulouse, France (2001) 290–297.
4. Ringel, M., Hirschberg, J.: Automated message prioritization: Making voicemail retrieval more efficient. In: Proc. Conf. on Human Factors in Computing Systems (Ext. Abstracts), Minneapolis, MN, USA (2002) 592–593.
5. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval **1** (1999) 67–88.
6. Lewis, D.D., Schapire, R.E., Callan, J.P., Papka, R.: Algorithms for linear text classifiers. In: Proc. 19th annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval. (1996) 298–306.
7. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47.
8. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow` (1996).
9. Teahan, W.J., Harper, D.J.: Using compression based language models for text categorization. In: Proc. Workshop on Language Modeling and Information Retrieval, Carnegie Mellon University, USA (2001) 83–88.
10. Peng, F., Schuurmans, D., Kaselj, V., Wang, S.: Automated authorship attribution with character level language models. In: Proc. 10th Conf. of European Chapter of Assoc. for Computational Linguistics, Budapest, Hungary (2003) 19–24.
11. Padmanabhan, M., Eide, E., Ramabhardan, G., Ramaswany, G., Bahl, L.: Speech recognition performance on a voicemail transcription task. In: Proc. IEEE ICASSP, Seattle, WA, USA (1998) 913–916.
12. Koumpis, K., Renals, S.: The role of prosody in a voicemail summarization system. In: Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ, USA (2001) 87–92.
13. Cordoba, R., Woodland, P.C., Gales, M.J.F.: Improving cross task performance using MMI training. In: Proc. IEEE ICASSP. Volume 1, Orlando, FL, USA (2002) 85–88.
14. Rosch, E.: Principles of categorization. In Rosch, E., Lloyd, B.B., (Eds.): Cognition and Categorization. Erlbaum, Hillsdale, NJ, USA (1978) 27–48.
15. Charlet, D.: Speaker indexing for retrieval of voicemail messages. In: Proc. IEEE ICASSP. Volume 1, Orlando, FL, USA (2002) 121–124.
16. Gotoh, Y., Renals, S.: Statistical language modelling. In: Renals, S., Grefenstette, G., (Eds.): Text and Speech Triggered Information Access. Springer-Verlag, Heidelberg, Germany (2003) 78–105.
17. Jelinek, F., Mercer, R.L., Bahl, L.R., Baker, J.K.: Perplexity – a measure of difficulty of speech recognition tasks. In: Proc. 94th Meeting Acoustical Society of America, Miami Beach, FL, USA (1977).
18. Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech and Language **13** (1999) 359–394.
19. Porter, M.: An algorithm for suffix stripping. Program **14** (1980) 130–137.