Multi-class Extractive Voicemail Summarization

Konstantinos Koumpis and Steve Renals

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St., Sheffield S1 4DP, UK
{k.koumpis,s.renals}@dcs.shef.ac.uk

Abstract

This paper is about a system that extracts principal content words from speech-recognized transcripts of voicemail messages and classifies them into proper names, telephone numbers, dates/times and 'other'. The short text summaries generated are suitable for mobile messaging applications. The system uses a set of classifiers to identify the summary words, with each word being identified by a vector of lexical and prosodic features. The features are selected using Parcel, an ROC-based algorithm. We visually compare the role of a large number of individual features and discuss effective ways to combine them. We finally evaluate their performance on manual and automatic transcriptions derived from two different speech recognition systems.

1. Introduction

Humans can recognise the gist of what was said rather than the precise word sequence. In addition, they can assimilate information faster through the eyes than the ears. Listening to a speech recording requires in general more effort than visually scanning its transcription because of the transient and temporal nature of audio. Audio recordings capture the richness of speech, yet it is not easy to directly browse the stored information. At the same time, transcribed spontaneous speech usually includes various kinds of redundant information.

In this paper we address the case of automatic generation of short text summaries of voicemail messages suitable for transmission as Short Message Service (SMS) text messages. Automatically produced text summaries from voicemail messages may serve multiple goals including rapid digestion of content, as well as indexing of messages with the intention of retrieving the original recordings or full transcriptions when more information is needed. Voicemail summarization has several features that differentiate it from conventional text summarization. Typical voicemail messages are short: the average duration of a voicemail message is 40s in the work reported here. The summaries are extremely terse, in this case designed to fit into a 140 character long text message and therefore coherence and document flow (style) are less important than informativeness. Only one speaker speaks at a time and due to the relatively short length of messages segmentation is not necessary (in contrast to spoken dialogues or broadcast news). Since the voicemail messages are transcribed by an automatic speech recognition (ASR) system, a significant word error rate (WER) must be assumed.

We have adopted a *word-extractive* approach [1] to voicemail summarization defining a summary as a set of content words extracted from the original message transcription. Although according to the above definition each word is treated independently, when humans listen to spoken utterances they too identify individual words prior to extracting a linguistic meaning from them. Given a spoken message $\mathcal{S}$, the extractive summarisation task can be framed as the mapping of each transcribed word into a predefined summary class:

$$\mathcal{S} : \mathrm{R}^{|\mathrm{V}|} \rightarrow \{0,1\}^{|\mathrm{C}|} \tag{1}$$

where $|\mathrm{V}|$ is the vocabulary size and $|\mathrm{C}|$ is the class-set size.

We have previously described and systematically evaluated a binary decision task [2] in which classifiers are trained to discriminate between 'summary words' and non-summary words. In this paper we deal with five predefined classes, four classes containing the words that carry principal content, namely proper names, telephone numbers, dates/times and 'other', and one class containing all redundant and irrelevant words. Hence, this multi-class summarisation task is expressed by (1) for $|\mathrm{C}| = 5$. Each word in the transcribed message is represented as a vector of lexical and raw prosodic features. The multi-class summarisation task apart from its applicability in structured query scenarios, can be used to enhance the visual listing of summaries, e.g. different classes shown in different colour, or associated with specific functionalities, e.g. telephone numbers with speed dialing.

A number of techniques have been proposed to extract key pieces of information from voicemail messages. Huang et al. [3] discussed three approaches to extract the identity and phone number of the caller: 200 hand-crafted rules; grammatical inference of subsequential transducers; and log-linear classifiers with 10 000 bigram and trigram features used as taggers. Jansche and Abney [4] proposed a phone number extractor based on a two-phase procedure that employs a hand-crafted component derived from empirical data distributions, followed by a decision tree that takes the length of a candidate phone number into account. The above techniques rely explicitly on lexical information and the best performing methods are based on hand-crafted rules.

The rest of this paper is organized as follows. An overview of the data and the annotation protocol we used in section 2 is followed by the feature comparison in section 3. The feature selection is described along with the summaries generation and evaluation in section 4. The paper is concluded in section 5.

2. Voicemail speech data

We have used the IBM Voicemail Corpus-Part I [5], distributed by the Linguistic Data Consortium (LDC). This corpus contains 1801 messages (14.6 hours, averaging about 90 words per message). We have two test sets: the 42 message development test set distributed with the corpus (referred to as test42) and a second 50 message test set provided by IBM (test50). The messages in test42 are rather short, averaging about 50 words per message, whereas the messages in test50 are closer to the training set average of 90 words per message.

As described in [6], we built a hybrid multi-layer perceptron (MLP) / hidden Markov model (HMM) recognizer with a combination of front-ends and multi-style trained language models. The essence of the hybrid approach is to train neural network classifiers to estimate the posterior probability of context independent phone classes, then to use these probabilities (converted into likelihoods by dividing with the priors) as inputs to a HMM decoder. During speech recognition training, we reserved the last 200 messages of the corpus as a validation set, resulting in a 1601 message training set. The average test set WERs were 41% on test42 and 44% on test50. We denote these transcriptions SR_1. In comparison to better

K. Koumpis is currently with Domain Dynamics Ltd.

	Training	*Validation*	*test42*	*test50*
Messages	800	200	42	50
Transcribed words	66 049	17 676	1 914	4 223
Total content words	20 555	5 302	561	820
Proper names	2 451	666	111	170
Tel. numbers	3 007	577	120	190
Dates and times	1 862	518	46	81
Other	13 235	3 541	284	379
Compression rate	31%	30%	29%	19%

Table 1: Voicemail content word annotation.

performing Gaussian-mixture model based systems, the main difference was that our system lacked any adaptation to characteristics of individual speakers, which are not directly applicable to the hybrid MLP/HMM approach. In order to estimate the effects of WER we obtained a second set of transcriptions (denoted SR_2) produced by the more complex HTK system developed for Switchboard and adapted to Voicemail corpus [7]. The WER for SR_2 was 31% for both test sets. The WER in voicemail corpus is not uniform, but is bursty within and across messages.

2.1. Annotation of summary words

As shown in Table 1 the first 800 messages of the Voicemail corpus were used as a summarization training set, and the last 200 used as a validation set. The transcriptions supplied with the Voicemail corpus include marking of named entities (NE), and we built on this using the following scheme:

1. Pre-annotated NEs were marked as targets, unless unmarked by later rules;
2. The first occurrences of the names of the speaker and recipient were always marked as targets; later repetitions were unmarked unless they resolved ambiguities;
3. Any words that explicitly determined the reason for calling including important dates/times and action items were marked;
4. Words in a stopword list with 54 entries were unmarked;
5. All annotation was performed using the human transcription only (no audio) so as not to introduce a bias towards acoustically prominent words.

The compression rate in our training, validation and testing material was in the range of 19% to 31%.

3. Feature comparison

The observation space comprised a total of 24 lexical and prosodic features listed in Table 2. Using them as inputs to a linear classifier we obtained the receiver operating characteristic (ROC) curves shown in Figure 1 for each of the four summary classes within the multi-class voicemail summarisation task. ROC curves are obtained by plotting sensitivity ($= \frac{\mathrm{TP}}{\mathrm{TP+FN}} =$ true positive rate) and [1 – specificity] ($= 1 - \frac{\mathrm{TN}}{\mathrm{TN+FP}} =$ false negative rate) to the vertical and horizontal axes, respectively, for various decision criteria of a classification task.

3.1. Proper names class

Proper names were discriminated very accurately by class specific NE matching ($\mathrm{ne}_{\mathrm{1(nam)}}$). However, the sensitivity was reduced dramatically when matching with the stemmed list ($\mathrm{ne}_{\mathrm{e(nam)}}$) was used. NE matching based on all types of NE ($\mathrm{ne}_{\mathrm{1(all)}}$) gave a good discrimination as the related list entries contain mostly proper names. This was less evident though with the increased confusion introduced by the stemmed list ($\mathrm{ne}_{\mathrm{2(all)}}$). Collection frequency too offered good discrimination with the variant based on stemmed words (cf_2) performing slightly but consistently better than cf_1. Word position (pos) had strong negative correlation with this summary class, indicating that proper names are mostly positioned at the beginning of voicemail transcriptions where the position features have low values. We also observed a low acoustic confidence for proper names. Regarding the prosodic features mean RMS energy (e), $\mathrm{F}_{\mathrm{0(ran)}}$, F_0 and duration (in descending order) gave useful discrimination. A weak correlation with following pauses (fp) was also observed.

Lexical Features
ac: acoustic confidence
cf_1: collection frequency of actual words
cf_2: collection frequency of stemmed words
$\mathrm{ne}_{\mathrm{1(all)}}$: matching of all actual NE words*
$\mathrm{ne}_{\mathrm{2(all)}}$: matching of all stemmed NE words*
$\mathrm{ne}_{\mathrm{1(nam)}}$: matching of proper names*
$\mathrm{ne}_{\mathrm{2(nam)}}$: matching of stemmed proper names*
$\mathrm{ne}_{\mathrm{1(tel)}}$: matching of telephone numbers*
$\mathrm{ne}_{\mathrm{2(tel)}}$: matching of stemmed telephone numbers*
$\mathrm{ne}_{\mathrm{1(d/t)}}$: matching of dates and times*
$\mathrm{ne}_{\mathrm{2(d/t)}}$: matching of stemmed dates and times*
$\mathrm{ne}_{\mathrm{1(oth)}}$: matching of other NE words*
$\mathrm{ne}_{\mathrm{2(oth)}}$: matching of stemmed other NE words*
pos: word position within message

Prosodic Features
dur_1: duration normalized by corpus
dur_2: duration normalized by message ROS
pp: preceding pause*
fp: succeeding pause*
e: mean RMS energy normalized by message
$\Delta\mathrm{F}_0$: delta of F_0 normalized by message
F_0: average F_0 normalized by message
$\mathrm{F}_{\mathrm{0(ran)}}$: F_0 range
$\mathrm{F}_{\mathrm{0(on)}}$: F_0 onset
$\mathrm{F}_{\mathrm{0(off)}}$: F_0 offset

Table 2: Lexical and prosodic features calculated for each word in the voicemail training, validation and test sets for the multi-class summarization task. The features marked with an asterisk (*) are represented by binary variables.

3.2. Telephone numbers class

Telephone numbers were discriminated very accurately by both class specific NE matching features ($\mathrm{ne}_{\mathrm{1(tel)}}$, $\mathrm{ne}_{\mathrm{2(tel)}}$), and in a lesser degree by the date and time specific NE matching features ($\mathrm{ne}_{\mathrm{1(d/t)}}$,$\mathrm{ne}_{\mathrm{2(d/t)}}$). This can be explained by the fact that these two classes share a large number of entries, namely digits. Word position (pos) offered a good discrimination as telephone numbers typically appear towards the end of a message. Collection frequency had an interesting correlation with this class. For words with low collection frequency the correlation was strongly negative, while the correlation was slightly positive for words with a collection frequency above the average. It is also notable that the telephone numbers class had the highest acoustic confidence among all summary classes. From the prosodic features only the durational ones proved to be correlated with telephone numbers. The rest of prosodic features did not offer any useful discrimination.

3.3. Dates and times class

Dates and times were discriminated adequately by the class specific NE matching features ($\mathrm{ne}_{\mathrm{1(d/t)}}$,$\mathrm{ne}_{\mathrm{2(d/t)}}$) in a similar fashion to telephone numbers. However, the telephone numbers specific NE matching ($\mathrm{ne}_{\mathrm{1(tel)}}$, $\mathrm{ne}_{\mathrm{2(tel)}}$) had a much lower sensitivity with dates and times class due to several irrelevant entries. Notably, $\mathrm{ne}_{\mathrm{1(all)}}$ offered better discrimination with respect to dates and times than $\mathrm{ne}_{\mathrm{1(tel)}}$ or $\mathrm{ne}_{\mathrm{2(tel)}}$. In contrast to telephone numbers, collection frequency features (cf_1,cf_2) correlated well with dates and times. The most important prosodic features were found to be fp, dur_1 (similar to dur_2), $\mathrm{F}_{\mathrm{0(off)}}$ and F_0 (in descending order). $\mathrm{F}_{\mathrm{0(off)}}$ had the largest correla-

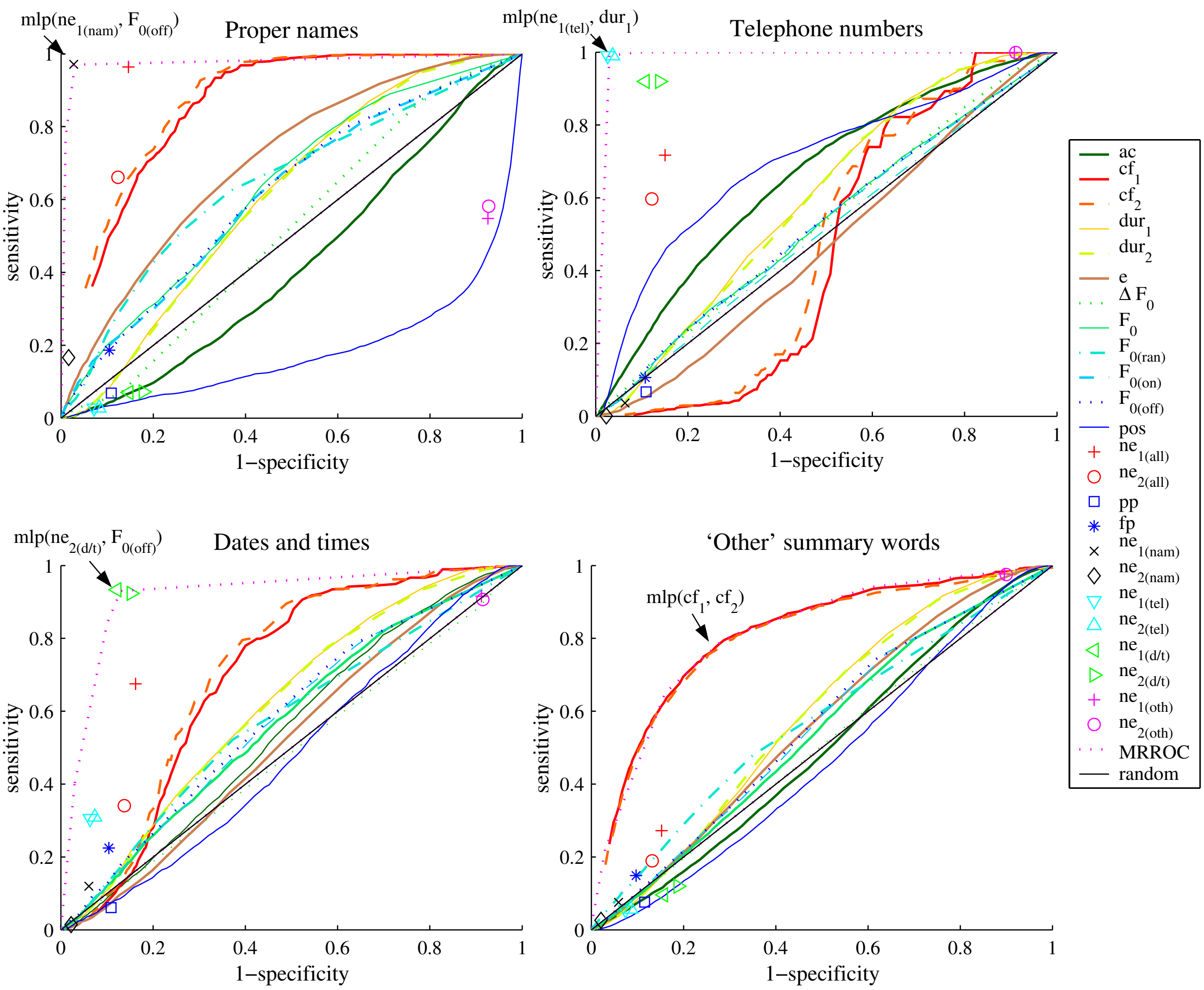

Figure 1: The ROC curves produced for the multi-class summarisation task using the individual features listed in Table 2 as inputs to a linear classifier with respect to the validation set. The MRROC curves produced by Parcel are also shown, along with the input features to MLPs at the MRROC vertex closer to the optimal classifier.

tion with dates and times class comparing with all other summary classes. Dates and times class was the only class with a negative correlation with energy features. The rest of prosodic features did not offer any useful discrimination.

3.4. Other class

The words belonging into the 'other' class proved to be the most difficult to discriminate using single features. The best correlating features with this general class were related to collection frequency (no difference between cf_1 and cf_2). As expected for this class, NE related features offered virtually no discrimination, with the slightly positive correlation explained by partial word matching with the NE lists. Among the prosodic features dur_1 (similar to dur_2), $\mathrm{F}_{0(\mathrm{rang})}$, mean RMS energy and $\mathrm{F}_{0(\mathrm{off})}$ offered some discrimination.

4. Feature selection for summarization

From the above analysis it is evident that distributions of most features vary considerably for different summary classes while the extent of overlap among classes is also significant. We used the data to guide us to an optimal subset of features Parcel framework [8]. Parcel does not select a single feature subset (or classifier), rather it selects as many subsets that are required to maximize performance at all operating points. We evaluated features using a sequential forward selection method and estimated the class probabilities using single layer network (SLN) and MLP classifiers with softmax output unit activation function. Parcel starts by estimating classifiers using single features only, and forms the maximum realizable ROC curve (MRROC). Those classifiers that are vertices of the MRROC are retained. If there are n total features, and a retained classifier uses a subset of k features, then $n - k$ new classifiers are generated, by adding each of the unused features to the feature set. The new classifiers are trained, the MRROC is updated and the process continues. The algorithm terminates when retraining any of the retained classifiers with an additional feature does not extend the MRROC.

The ROC curves produced by Parcel for the multi-class summarisation task on the validation set using the SLN and the MLP classifiers were almost identical for the proper names, telephone numbers and dates and times classes. For the 'other' class the accuracy was complementary, but still comparable. The vertices of the MRROC curves with the shortest Euclidean distance from the optimal classifier denoted by the (0,1) point in the ROC space, i.e. the one with the best trade-off between

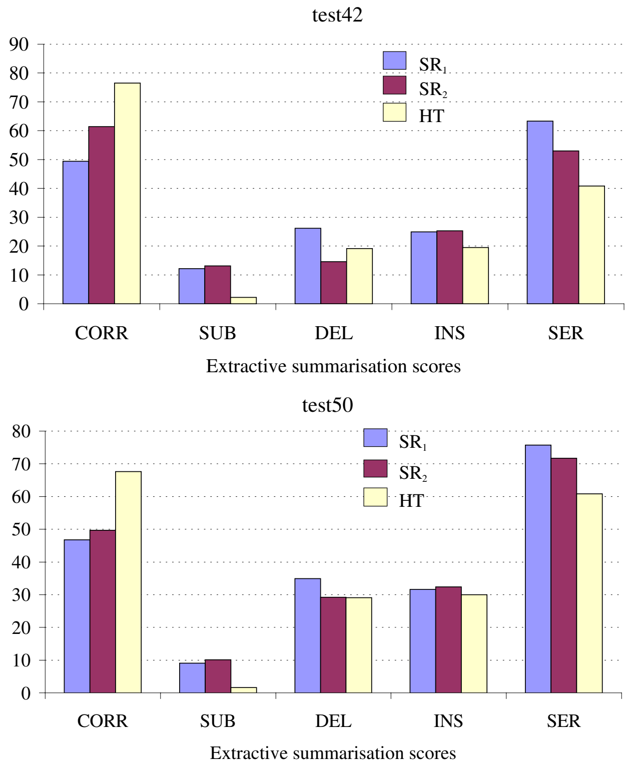

Figure 2: Extractive summarization scores on test42 and test50 for SR_1, SR_2 and human transcription (HT) input, respectively.

sensitivity and [1 – specificity], are also shown in Figure 1. These were produced by an MLP with 20 hidden units and the following input features: $\mathrm{ne}_{1(\mathrm{nam})}$ and $\mathrm{F}_{0(\mathrm{off})}$ for proper names; $\mathrm{ne}_{1(\mathrm{tel})}$ and dur_1 for telephone numbers; $\mathrm{ne}_{2(\mathrm{d/t})}$ and $\mathrm{F}_{0(\mathrm{off})}$ for dates and times; and cf_1 and cf_2 for the 'other' words class. The fact that we used class specific NE matching features caused some overfitting in the validation set and given that we required a minimum difference of 5% in the area under the MR-ROC for the Parcel to continue, lead to relatively small feature subsets containing 2-3 features each.

The combination of different classifier outputs within the multi-class summarisation task is not straightforward. One way is to combine the areas under the MRROCs of different classes into a single weighted sum where the weight of each class is proportional to the class's prior in the training set. However, this is in general non-intuitive and computationally expensive. Furthermore, some class priors estimated over the training set can be totally irrelevant to the class distributions of some messages. For instance, consider messages where there are repetitions of the same content e.g. telephone numbers and thus many words belonging into a certain class with a low prior. If the classifiers for the classes with high priors are applied according to these priors, they might not leave enough room to extract words from the remaining classes given the restrictions imposed by the compression rate.

Due to the above reasons we took a different approach in which we applied sequentially one classification system for each of the summary classes. We started with the one having the largest area under the MRROC and continued with the remaining three in descending order. The restriction was that a classification system with a smaller area under the MRROC than the ones previously applied could not remove already selected summary words.

The summaries produced with the above procedure were evaluated using a slot error rate (SER) measure that treats substitution errors (correct classification, wrong transcription), insertion errors (false positives) and deletion errors (false negatives) equally. The scores derived for test42 and test50 are given in Figure 2 where CORR refers to correct word transcription and classification. The SER on test42 was 63%, 53% and 41% for SR_1, SR_2 and HT outputs, respectively. Regarding test50 the SER scores were 76%, 72% and 61% for SR_1, SR_2 and HT outputs, respectively. Although the overall summary word detection rate was not improved in comparison with the binary task [2], names, dates and times, and numbers could be detected with higher accuracy.

5. Conclusion

We have presented a system for the multi-class word-extractive summarization of voicemail based on the selection of lexical and prosodic features. From the ROC analysis it became evident that distributions of most features vary considerably for different summary classes while the extent of overlap among classes is also significant. Prosodic features played a significant role in extracting proper names, telephone numbers and dates/times. The 'other' summary words were discriminated almost exclusively by collection frequency. We generated summaries by applying sequentially the classification system with the best tradeoff for each summary class. The evaluation showed that although the overall summary word detection rate was not improved in comparison with the binary task, names, dates and times, and numbers could be detected with higher accuracy.

6. Acknowledgements

We thank Mark Gales (Cambridge University) for providing us with the HTK transcriptions of the voicemail test sets. This work was supported by EPSRC ROPA award GR/R23954.

7. References

[1] K. Koumpis, S. Renals, and M. Niranjan. Extractive summarization of voicemail using lexical and prosodic feature subset selection. In *Proc. Eurospeech*, pages 2377–2380, Aalborg, Denmark, 2001.

[2] K. Koumpis and S. Renals. Evaluation of extractive voicemail summarization. In *Proc. ISCA Workshop on Multilingual Spoken Document Retrieval*, Hong Kong, China, 2003.

[3] J. Huang, G. Zweig, and M. Padmanabhan. Information extraction from voicemail. In *39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France., 2001.

[4] M. Jansche and S. Abney. Information extraction from voicemail transcripts. In *Proc. Conference on Empirical Methods in NLP*, Philadelphia, PA, USA, 2002.

[5] M. Padmanabhan, E. Eide, G. Ramabhardan, G. Ramaswany, and L. Bahl. Speech recognition performance on a voicemail transcription task. In *Proc. IEEE ICASSP*, pages 913–916, Seattle, WA, USA, 1998.

[6] K. Koumpis and S. Renals. The role of prosody in a voicemail summarization system. In *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 87–92, Red Bank, NJ, USA, 2001.

[7] R. Cordoba, P. C. Woodland, and M. J. F. Gales. Improving cross task performance using MMI training. In *Proc. IEEE ICASSP*, volume 1, pages 85–88, Orlando, FL, USA, 2002.

[8] M. Scott, M. Niranjan, and R. Prager. Parcel: Feature subset selection in variable cost domains. Technical report, CUED TR-323, ftp://svr-ftp.eng.cam.ac.uk/pub/reports, Cambridge, UK, 1998.