AUTOMATIC BROADCAST NEWS SUMMARIZATION VIA RANK CLASSIFIERS AND CROWDSOURCED ANNOTATION

Srinivas Parthasarathy and Taufiq Hasan

Research and Technology Center
Robert Bosch LLC
taufiq.hasan@us.bosch.com

ABSTRACT

Extractive speech summarization methods generally operate as a binary classifier deciding if a sentence belongs to the summary or not. However, it is well known that even human annotators do not agree on selecting most summary sentences. In this paper, we take a probabilistic view of the summarization ground-truth and assume that more frequently selected sentences by annotators are of higher importance. Using a large summary data-set obtained through crowdsourcing, we empirically show that sentence selection frequency is inversely related to its summarization rank. Consequently, we model the relative importance between sentences using a rank-based classifier. Additionally, we utilize an extended paralinguistic feature set that has not been previously used for speech summarization. Lexical and structural features are also included. Support Vector Machine (SVM) is used as the baseline binary classifier and rank classifier. Experimental evaluations show that the proposed approach outperforms traditional binary classifiers with respect to various ROUGE summarization metrics for different summarization compression ratios (CR).

Index Terms— Spoken document summarization, paralinguistic features, crowdsourcing

1. INTRODUCTION

Automatic speech summarization aims at identifying the most important and relevant information from an audio signal and generate a compressed version of the original spoken document [1, 2]. By nature, speech summarization is generally considered an extractive process [3, 4, 5, 6] since the summary must be generated through combining partitions of the input signal. In contrast to text summarization, summarizing a speech signal constitutes of unique challenges, such as, errors in automatic speech recognition (ASR), disfluencies and redundancy in spontaneous speech, difficulty in sentence endpoint detection, presence of background noise, etc.

In the past, research on speech summarization approaches utilized various methods including document structure, linguistic and prosodic information, and significance measures in identifying representative sentences [1]. Recently, supervised machine learning based summarization methods have been widely studied with promising results in various scenarios [7, 3, 4, 8]. Most supervised methods view extractive summarization as a binary classification task. Reference spoken documents are transcribed and human annotators are instructed to select sentences that most effectively summarize the document. Various acoustic, prosodic, lexical and structural features [3] are extracted from each sentence and the classifier is trained to distinguish between the two classes: summary and non-summary. During evaluation, sentences from the unseen document are rank ordered based on their posterior probability of belonging to the summary class.

One of the major issues in speech summarization is the problem of low inter-annotator agreement. Most research work utilize human generated summaries prepared by a few annotators. The annotators select sentences or sentence-like units (SU) as representative of the spoken document which are used as gold-standard summaries. However, it is generally known that agreement among annotators is quite low [9] with average Kappa coefficient between $0.2 - 0.3$. Generally, multiple annotators agree on selecting a few sentences in the summary while many sentences are only selected by a single annotator [2]. This is expected since a given document can be summarized in different ways and it is extremely difficult to evaluate their relative quality [10]. The fact that multiple good summaries can exist is also well known and considered in calculating the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluation metric [11].

In this paper, we assume that extractive summarization annotations are inherently fuzzy and a clear distinction between summary and non-summary is not optimal [12, 4, 6]. We hypothesize that some sentences are more important or relevant than others, which are usually the few sentences the annotators agree on selecting. It is difficult to quantify the relative importance among sentences unless a large number of summary ground-truths are available for each document. For this purpose, we utilize crowdsourcing to annotate a large number of broadcast news stories. To the best of our knowledge, a crowdsourcing study on this scale has not been done for broadcast speech summarization in the past.

There are a few alternatives to the binary classification scheme for extractive summarization, including rank classi-

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

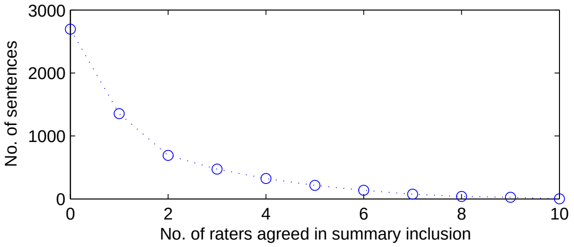

Fig. 1. Distribution of sentences in different annotator agreement level. Most sentences are not included in the summary at all. Very few are selected by all annotators.

fiers [12] and regression [6]. In the former case, a pair-wise importance rank is learned from the features obtained from each sentence. During the evaluation phase, the sentences from the unknown story can be directly rank ordered. In the regression approach [6], the authors presented different ways of using a real-valued label to each sentence representing their importance instead of binary ones. A support vector regression model is used to learn these importance values as ground-truth and estimate them during evaluation. In this work, we utilize both binary and rank classifiers using the MTurk annotations. We assume that, sentence importance is proportional to the number of rater agreements and use this information to train our classifiers.

2. CORPUS AND ANNOTATIONS

In this work, we used the RT-03 MDE Training Data Speech and annotations collected by the Linguistic Data Consortium (LDC) [13]. There are approximately 20 hours of Broadcast News and over 40 hours of Conversational Telephone Speech contained in this corpus. The broadcast news data is a subset of the 1997 English Broadcast News Speech (HUB4) corpus. Sentence level annotations and speaker information are also available along with the transcripts. In this data-set, each audio file correspond to a single news story. We have selected 90 such news stories for the evaluation of our speech summarization system.

2.1. Annotation using Mechanical Turk

In order to obtain gold-standard summarization prepared by human annotators, we utilize the croudsourcing service Mechanical Turk (MTurk) provided by Amazon. The news stories were published as Human Intelligence Tasks (HITS). For a given story, each of it's sentence is displayed with a checkbox on the left side. The assessors (workers) are instructed to read the story carefully and select $10-15\%$ sentences that best summarize the article. We provide some flexibility in the selection amount since not all sentences are of equal length. Each story was annotated by 10 independent assessors. In order to verify that the workers read the sentences carefully [14], we include a mandatory sentence in a random position within the article. This sentence lets the worker know that it must be selected to be approved. By verifying that the annotators selected the mandatory sentence, we can identify if the task was performed carefully. In addition, we restrict the task for only the workers who have an HIT approval rate of 98% and have performed at least 5000 tasks in the past. These criteria ensures that only the serious and dedicated workers are allowed to perform our summarization HITs.

Table 1. Level of agreement among the ten annotators providing 10% summaries. Annotators are not the same across stories. Values are averaged across all stories and annotators.

ROUGE-1	ROUGE-2	ROUGE-L	Kappa (κ)
0.54733	0.38710	0.53411	0.2057

2.2. Inter-annotator agreement analysis

We perform a series of analysis to evaluate the agreement among the raters. Firstly, we use the ROUGE scores within the reference summaries as performed in [4]. For each story, one assessor's summary is used as reference while the remaining 9 are assumed to be automatic summaries. This is repeated for all the assessors and the average ROUGE metric is obtained across all the stories. Using the recall based ROUGE metric and Porter stemmer, ROUGE-1 (unigram match), ROUGE-2 (bi-gram match) and ROUGE-L (longest common sequence match) values are computed [11]. Averaged Fleiss' Kappa coefficient for inter-rater agreement [15] is also computed across the stories. These metrics are shown in Table 1. Also, in Fig. 1, the distribution of sentences in different agreement levels is shown. Here, we observe that majority of the sentences are not selected for summary by any annotator while only a few are selected by many.

Observing the values in Table 1, it is evident that the agreement between annotators is not very strong, which is consistent with previous findings [9]. In our view, this low inter-annotator agreement is firstly due to the nature of the summarization task, since a document can be summarized using different combination of sentences. Secondly, it is due to the issue of noisy annotations obtained through MTurk itself. However, with sufficient annotations per story, we can learn the relative importance between sentences rather than distinguishing them in two classes.

2.3. A probabilistic view of summary annotations

It should be expected that an important sentence in a news story will be included in the summary by multiple annotators. Thus, the importance of a sentence is proportional to its probability of selection. We are interested to observe the distribution of this probability with sentence rank. Assume a single story s_i contains N_i sentences: $\{t_{i1}, t_{i2}\cdots t_{iN_i}\}$, each of which were selected $\{n_{i1}, n_{i2}\cdots n_{iN_i}\}$ times by annotators,

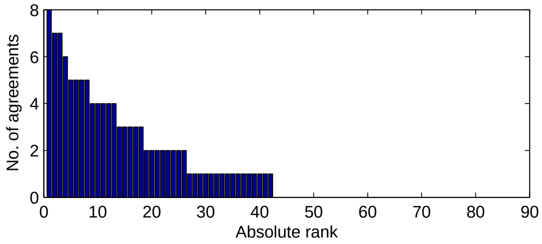

Fig. 2. Number of annotator agreements vs rank of sentences for a single news story. Top ranked sentences are selected by most raters while majority are selected by none or a few.

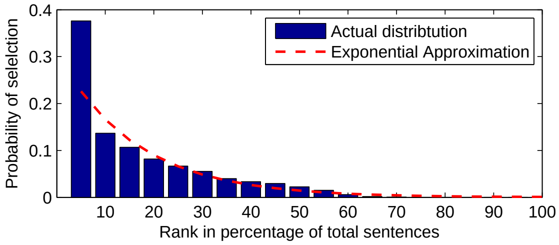

Fig. 3. Probability distribution of sentence rank in percentage of total sentences (in bins of 5%). Sentence selection probability decreases exponentially with increasing rank.

respectively. If the sentences t_{ij} are rank ordered according to their corresponding n_{ij} values, we obtain a frequency table showing the distribution of rank vs. agreement as shown in Fig.2. Lets assume that the rank of t_{ij} is given by an integer $r_{ij} \in [1, N_i] \cap Z$. Here, we observe that a few sentences are selected by many raters while others only by a few. There is a drastic drop in sentence selection frequency with increasing rank, resembling Zipf's law.

If we assume that rank is an inherent property of a sentence for a given story, we can estimate a probability density function of sentence rank. For a single story, the probability of selecting sentence t_{ij} would be $p(t_{ij}) = n_{ij}/\sum_j n_{ij}$. In order to aggregate the annotation data from all stories s_i of different lengths N_i, we define a relative rank $\rho_{ij} = r_{ij}/N_i \in (0, 1]$. Next, we compute n_{ij} and corresponding ρ_{ij} values for all stories. Accumulating the values of n_{ij} in ρ_{ij} ranges of $(0-5)\%, .., (5-100)\%$, we obtain a frequency table, which in turn can be converted to a probability density function by dividing each bin value by $\sum_i \sum_j n_{ij}$. The resulting distribution is shown in Fig. 3. Here, we observe a clear exponential decay of sentence selection probability (or importance) with increasing rank. For example, the sentences in top 5% rank is twice as more important than sentences in $5-10\%$ rank, with corresponding selection probabilities of 0.376 and 0.136, respectively.

3. SUMMARIZATION SYSTEM COMPONENTS

3.1. Segmentation

Previous studies have shown that sentence units are ideal units for summarization. The transcripts for the LDC corpora provide sentence level annotations with start and end times. In our work, we utilize these time-stamps and corresponding transcripts for audio segmentation and feature extraction.

3.2. Features

3.2.1. Paralinguistic features

We extract various acoustic and prosodic features from each speech sentence. This feature set was introduced by Schuller et al. [16] for the Speaker State Challenge at Interspeech 2011. At first, the frame level features describing acoustic properties, known as *Low level Descriptors(LLD)*, are extracted. These are described in Table 3. Next, for each sentence in the corpus a global statistic of these features is calculated, yielding *high level features (HLF)*. The global statistic produces a single value for the entire sentence which is independent of the sentence length. The total feature set consists of 4368 dimensions. Due the high dimensionality of the feature set a correlation based feature selection is performed. The reduced feature set consists of 110 dimensional paralinguistic features.

Table 3. Low-level descriptors (LLD)

Energy features (dimension 4)
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
Speech spectral features (dimension 50)
RASTA-style filter auditory spectrum, bands $1-26$ (08 kHz)
Mel Frequency Cepstral Coefficients (MFCC) $1-12$
Spectral energy $25-650$ Hz, $1-4$ kHz
Spectral Roll Off Point $0.25, 0.50, 0.75, 0.90$
Spctral Flux, Entropy, Variance, Skewness, Kurtosis, Slope
Voice related (dimension 5)
F0, Probability of voicing
Jitter (local & delta), Shimmer (local)

3.2.2. Lexical features

The following lexical features are extracted from each sentence: i) number of words, ii) number of Named Entities (NE), iii) number of stop-words, iv) sentiment polarity, v) TF-IDF (Term frequency - Inverse Document Frequency) vector, and v) bi-gram language model scores. The Natural Language Toolkit (NLTK) library [17] is utilized for sentiment analysis and NE extraction. The IDF values were computed using a dictionary extracted from Wikipedia and the Google N-gram data [18]. TF-IDF features are extracted from each word from

Table 2. Experimental results comparing the baseline binary SVM and the proposed rank-SVM classifiers with respect to ROUGE-1, ROUGE-2 and ROUGE-L metrics.

Metric	ROUGE-1			ROUGE-2			ROUGE-L		
Compression	5%	10%	15%	5%	10%	15%	5%	10%	15%
Binary SVM baseline	0.28258	0.45911	0.58506	0.18473	0.30051	0.39833	0.27147	0.44385	0.56900
Rank-SVM-1	0.32538	0.53611	0.66262	0.21632	0.37027	0.48196	0.31325	0.52000	0.64797
Rank-SVM-2	0.33591	0.54362	0.67031	0.22628	0.37972	0.49236	0.32347	0.52825	0.65561

a given sentence. The bi-gram language model was trained on the Brown and ABC corpus provided with NLTK. For this feature, the log-likelihood score of each bi-gram in a sentence is computed. For both TF-IDF and bi-gram scores, the minimum, maximum, average and summation of the values obtained from a sentence is used features, yielding a total dimension of 12.

3.2.3. *Structural features*

Past studies have shown that broadcast news summarization can be facilitated by features extracted from the structure of the stories [19]. In this work, we considered the following structural features: i) duration of the sentence, ii) duration of the sentence preceding and iii) following the current sentence and iv) position of the current sentence within the story.

3.3. Classifiers

Our analysis of the corpus in the previous sections show that sentences in the summary can be ranked based on the agreement of annotators, or probability of selection. Accordingly, we implement rank learning techniques based on the importance of each sentence. Many such techniques have been explored for information retrieval [20, 21, 22]. In this study we use a *pair-wise* approach. Given the sentences in a story $\{s_1, s_2, ..., s_n\}$, the objective of the classifier is to learn the preference relationship between pairs of instances and produce the order of sentences $s_1 > s_2 > s_3.... > s_n$. A significant advantage of rank classifiers for the summarization process is that since they learn the relative importance between different sentences, a single classifier can be used to produce summaries of varying compression ratios. We implement the Rank-SVM introduced by Joachims [20] in this work. We also use SVM as a binary classifier for comparison.

4. EXPERIMENTS AND RESULTS

In order to train the binary classifier, we use the top 10% ranked sentences according to the annotator agreement counts. These are used as *summary* while the remaining sentence are considered *non-summary*. We train the rank-SVM classifier in two different ways. For Rank-SVM-1, we use pairs of sentences that have a difference of importance by at least 5 rater agreements. However, both sentences may actually belong to the summary according to some annotators. For Rank-SVM-2, we use pairs of sentences so that one of them is in the top 10% rank while the other one is not. This approach is similar to [12]. The difference between the two approaches are subtle, but both are based on *learning to rank* principle.

To generate the summary, we fix the output length to 5%, 10% and 15% of the total number of sentences. The top ranked sentences obtained from the classifiers are included in the summary. Using the 90 broadcast news stories, we perform a 3-fold cross validation experiment. In each fold, the algorithms are trained on 60 sentences and evaluated on the remaining 30. The ROUGE-1, ROUGE-2 and ROUGE-L metrics are used to evaluate the system summaries with respect to all 10 reference (annotated) summaries. The results are presented in Table 2.

From Table 2, we observe that the rank-SVM classifiers significantly outperforms the binary SVM classifier. This is true for all three %CR values across different ROUGE metrics. A relative improvement of 20% is achieved in ROUGE-2 metric for 15% CR using Rank-SVM-2 approach, which is significant considering that both classifiers are trained using the top 10% sentences. Performance of Rank-SVM-1 and Rank-SVM-2 seems to be very similar. This demonstrate the benefit of learning to rank approach in contrast to binary classification, as in the former approach only the relative importance is being learnt.

5. CONCLUSIONS

In this work, we have proposed a probabilistic view in extractive summarization labels and utilized a rank SVM classifier for selecting summary sentences. In contrast to traditional view, where sentences are either considered in summary vs. non-summary class, the proposed method learns the relative importance between sentences from a large summary annotation obtained through crowdsourcing. The evaluation results demonstrate that the proposed approach yields superior summarization performance in with respect to ROUGE metrics in different compression ratios.

6. ACKNOWLEDGEMENTS

We thank Dr. Yang Liu (The University of Texas at Dallas) and Dr. Lin Zhao for their valuable discussions.

7. REFERENCES

[1] Inderjeet Mani and Mark T Maybury, *Advances in automatic text summarization*, vol. 293, MIT Press, 1999.

[2] Ani Nenkova, Sameer Maskey, and Yang Liu, "Automatic summarization," in *Proc. ACL: Tutorial Abstracts*. ACL, 2011, p. 3.

[3] Sameer Maskey and Julia Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization.," in *Proc. Interspeech*, 2005, pp. 621–624.

[4] Berlin Chen, Shih-Hsiang Lin, Yu-Mei Chang, and Jia-Wen Liu, "Extractive speech summarization using evaluation metric-related training criteria," *Information Processing & Management*, vol. 49, no. 1, pp. 1–12, 2013.

[5] Fei Liu and Yang Liu, "Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 21, no. 7, pp. 1469–1480, 2013.

[6] Shasha Xie and Yang Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, vol. 24, no. 3, pp. 495–514, 2010.

[7] Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey, "From text to speech summarization.," in *Proc. IEEE ICASSP*, 2005, pp. 997–1000.

[8] Shih-Hsiang Lin, Berlin Chen, and Hsin-Min Wang, "A comparative study of probabilistic ranking models for chinese spoken document summarization," *ACM Trans. on Asian Language Information Process.*, vol. 8, no. 1, pp. 3, 2009.

[9] Fei Liu and Yang Liu, "What are meeting summaries?: an analysis of human extractive summaries in meeting corpus," in *Proc. ACL SIGdial*. ACL, 2008, pp. 80–83.

[10] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," in *Proc. SIGIR*. ACM, 1999, pp. 121–128.

[11] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.

[12] Shih-Hsiang Lin, Yu-Mei Chang, Jia-Wen Liu, and Berlin Chen, "Leveraging evaluation metric-related training criteria for speech summarization," in *Proc. IEEE ICASSP*. IEEE, 2010, pp. 5314–5317.

[13] Strassel Stephanie, Christopher Walker, and Haejoong Lee, "RT-03 MDE Training Data Speech LDC2004S08," [Online] `https://catalog.ldc.upenn.edu/LDC2004S08`, 2004.

[14] Aniket Kittur, Ed H Chi, and Bongwon Suh, "Crowdsourcing user studies with mechanical turk," in *Proc. SIGCHI*. ACM, 2008, pp. 453–456.

[15] Joseph L Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, no. 5, pp. 378, 1971.

[16] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski, "The interspeech 2011 speaker state challenge.," in *Proc. Interspeech*. ISCA, 2011, pp. 3201–3204.

[17] Steven Bird, "Nltk: the natural language toolkit," in *Proc. ACL COLING*. ACL, 2006, pp. 69–72.

[18] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al., "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, pp. 176–182, 2011.

[19] Sameer Maskey and Julia Hirschberg, "Automatic summarization of broadcast news using structural features," in *Proc. Interspeech*, 2003.

[20] Thorsten Joachims, "Optimizing search engines using clickthrough data," in *Proc. ACM SIGKDD*. ACM, 2002, pp. 133–142.

[21] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, "Learning to rank using gradient descent," in *Proc. ACM ICML*. ACM, 2005, pp. 89–96.

[22] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, "Learning to rank: from pairwise approach to listwise approach," in *Proc. ACM ICML*. ACM, 2007, pp. 129–136.