TOPIC IDENTIFICATION BASED EXTRINSIC EVALUATION OF SUMMARIZATION TECHNIQUES APPLIED TO CONVERSATIONAL SPEECH

David Harwath and Timothy J. Hazen

MIT Lincoln Laboratory, Lexington, Massachusetts, USA

ABSTRACT

Document summarization algorithms are most commonly evaluated according to the intrinsic quality of the summaries they produce. An alternate approach is to examine the extrinsic utility of a summary, measured by the ability of the summary to aid a human in the completion of a specific task. In this paper, we use topic identification as a proxy for relevancy determination in the context of an information retrieval task, and a summary is deemed effective if it enables a user to determine the topical content of a retrieved document. We utilize Amazon's Mechanical Turk service to perform a large-scale human study contrasting four different summarization systems applied to conversational speech from the Fisher Corpus. We show that these results appear to be correlated with the performance of an automated topic identification system, and argue that this automated system can act as a low-cost proxy for a human evaluation during the development stages of a summarization system.

Index Terms— Document Summarization, Topic Modeling

1. INTRODUCTION

Automatic document summarization algorithms are often evaluated by comparing the summaries they generate against summaries written by human experts. This is known as an *intrinsic* evaluation, and the most ubiquitious metric used for this paradigm is the ROUGE score [1]. While aspiring to generate human-quality summarization is a worthy goal, *extrinsic* metrics which measure the utility of a summarization method for achieving a specific goal are more appropriate for many tasks [2, 3].

In this paper, we consider the example of an information retrieval (IR) system. IR systems often display short summaries of retrieved documents in order to assist users in selecting documents relevant to their query or task. The evaluation metric for this use case should measure the user's ability to determine a document's relevancy based solely on reading a summary. In our experiments, we use topic identification (topic ID) performance as a proxy for relevancy determination. For this approach, documents in an evaluation corpus need only be labeled by topic, and do not require the more expensive creation of human-generated summaries that ROUGE requires. Topic labeled corpora for text and speech such as the TREC TDT corpora [4] and the Fisher Corpus [5] already exist.

We have conducted a large scale study of this extrinsic evaluation paradigm with Amazon's Mechanical Turk crowdsourcing service. Mechanical Turk has previously been employed in performing extrinsic evaluations of machine translations systems [6], and its use for the extrinsic evaluation of summarization systems has been suggested [7]. Our experiments compare various summarization techniques present in the literature, as well as novel techniques which we present in this paper.

While crowdsourcing is an inexpensive and efficient alternative to employing experts, automated evaluation tools are preferred by researchers for iterative testing during the development of new summarization algorithms. The results of our study imply that automatic topic ID performance is correlated with human topic ID performance on automatically generated summaries. This suggests that evaluations can utilize low cost automatic topic ID systems to simulate real users during the research, development and optimization of summarization techniques, particularly for IR applications that require summaries to convey topical relevancy to the user.

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. CORPUS OF EXPERIMENTAL DATA

The experiments in this paper all use a set of 1374 conversations extracted from the Fisher Corpus [5]. This corpus consists of audio from 10-minute-long telephone conversations between two people. Before the start of each conversation, the participants were prompted to discuss a specific topic. Data was collected from a set of 40 different prompted topics including relatively distinct topics (e.g. "Pets", "Movies", "Hobbies", etc.) as well as topics covering similar subject areas (e.g. "Family", "Life Partners", "Family Values"). The topical content in the data is generally dominated by discussion of these 40 prompted topics, however off-topic discussions can also be routinely observed in the data. Our experiments use only the human-generated text transcripts of the Fisher conversations.

3. SYSTEM DESCRIPTIONS

For our study, we created four different summarization systems based partially on existing approaches present in the literature, and partially on novel techniques which we developed for this study. Each of these systems is a single document summarization system which takes as input a single Fisher conversation and outputs an extractive summary for that document. Three of the systems perform utterance extraction, and the fourth extracts individual signature words. We describe each system along with the supporting models and preprocessing procedure below.

3.1. Probabilistic Latent Semantic Analysis

To support the modeling in various components of our system, we utilize a latent topic model. Probabilistic Latent Semantic Analysis (PLSA) [8] is a probabilistic model which learns the relationship between the space of observed words and a latent topic space. We

use the PLSA implementation described in [9], although other latent topic models such as Latent Dirichlet Allocation [10] could be substituted.

PLSA models a collection of documents $D = \{d_1, \ldots, d_{N_D}\}$, in which each document is represented as a sequence of words $d_i = \{w_1, \ldots, w_{N_{d_i}}\}$. A probability model is learned for observing word w in document d through a set of latent variables, $Z = \{z_1, \ldots, z_{N_z}\}$:

$$P(w|d) = \sum_{\forall z \in Z} P(w|z)P(z|d) \quad (1)$$

The set of latent topics, Z, is learned in an unsupervised fashion using the EM algorithm. Our PLSA model learned 40 latent topics over the full collection of 1374 Fisher documents.

3.2. Preprocessing

To aid in the creation of readable, relevant summaries of consistent length, each conversation undergoes a series of preprocessing stages.

Windowing: The conversation is processed as a stream of words with no regard to pauses or speaker turns. A sliding window is used to extract a sequence of frames from the word stream. Each of the utterance extracting systems uses a window shift of 3 words, but differ in their window length. For a window capturing n words, the i^{th} frame is given by:

$$f_i = \{w_{1+3*i}, \ldots, w_{1+3*i+n}\} \quad (2)$$

Pruning: A posterior log likelihood ratio score for each latent topic is computed for each frame:

$$s_{f_i,z_j} = \sum_{w \in f_i} c_{w,f_i} \log \frac{P(w|z_j)}{P(w|\bar{z}_j)} \quad (3)$$

Here, c_{w,f_i} represents the number of times word w appears in frame f_i, and $P(w|\bar{z}_j)$ is a conditional likelihood estimate for w over all $z \neq z_j$. A latent topic score vector associated with f_i is produced by concatenating these latent topics scores as follows:

$$\vec{s}_i = [s_{f_i,z_1}, \ldots, s_{f_i,z_{N_z}}]^T. \quad (4)$$

An $\vec{s}$ vector is computed for each frame in the document, and any frames with a topical score vector whose magnitude falls below a threshold are discarded. In practice, this process pruned approximately 35% to 40% of the frames from each document, which typically corresponded to non-topical or disfluent speech.

Segmentation: After pruning, the remaining sequence of frames is segmented using a minimum cut algorithm as described in [11]. For the segmentation procedure, the similarity between frames f_i and f_j is calculated using the cosine similarity measure between their corresponding vectors of topical scores, $\vec{s}_i$ and $\vec{s}_j$. The result of the procedure is a set of N_{seg} segments, where the i^{th} segment is a set of consecutive, topically similar frames $S_i = \{f_j, \ldots, f_k\}$. The segmentation algorithm is allowed to find a large number of segments, which are then clustered together in the next step.

Clustering: A set of frame clusters are initialized such that $C_i = S_i$, and an aggregate topical score vector for each cluster is calculated as:

$$\vec{s}_{C_i} = \sum_{s_j \in C_i} \vec{s}_j \quad (5)$$

The similarity between two clusters, C_i and C_j, is then calculated using a standard cosine similarity measure. Clusters are greedily merged if this similarity exceeds 0.5, and merging continues until there are no more cluster pairs with a similarity exceeding this threshold.

Cluster Selection: A single cluster representing the dominant topical theme of the document is chosen for the basis of the summarization, and the remaining clusters are discarded. To perform this step, a topical dominance score $\mathrm{Dom}(C_i)$ is calculated for each cluster:

$$\mathrm{Dom}(C_i) = N_{f,i} \frac{\sum_{\forall z} P(z|C_i)\mathcal{Q}(z)}{1 + H(Z|C_i)} \quad (6)$$

$N_{f,i}$ is the number of frames in C_i, which helps to favor large clusters. $H(Z|C_i)$ is the conditional entropy of the distribution over topics given the cluster. Since this term appears in the denominator, it will boost the dominance score for topically coherent clusters, i.e. clusters with a low conditional entropy. $\mathcal{Q}(z)$ is the quality score of topic z over the entire document collection, as derived in [9]. The quality score of a topic is given by

$$\mathcal{Q}(z) = P(z) * \mathcal{P}_{Z \to D}(z), \quad (7)$$

and is governed by $P(z)$, the portion of the corpus z represents, as well as its $Z \to D$ purity measure, given by:

$$\mathcal{P}_{Z \to D}(z) = \exp\left(\frac{\sum_{\forall d} P(z|d) \log P(z|d)}{\sum_{\forall d} P(z|d)}\right) \quad (8)$$

A topic z will have a high purity measure if it tends to dominate the documents in which it appears, and a smaller score if the topic is weakly spread across many documents. In practice, selecting a single dominant cluster has the effect of removing smaller portions of a conversation during which the participants stray off topic. Only frames appearing in the dominant cluster are provided to the final extractive summarization system.

3.3. Direct Modeling with LexRank Extraction of 3 Best Frames

The first system uses the weighted LexRank algorithm described in [12]. LexRank constructs a graphical representation of a document, where node i represents frame f_i, and the edges contain the pairwise similarity measures between nodes. Edge weights are calculated using the standard cosine similarity measure between TF-IDF weighted vectors of word counts for each frame.

The LexRank algorithm then ranks the graph nodes in terms of their centrality, i.e. the "most connected" nodes are ranked highest. To summarize the document, the top 3 frames are extracted. As in [12], a Maximum Marginal Relevance (MMR) based reranking scheme is used, which helps prevent overlapping frames from being extracted together. This system uses a window size of 15 words, resulting in a constant summary length of 45 words. Once the summary frames are extracted, speaker turn markings are inserted back into the text. Below is an example summary generated by system 1 for a conversation on the "Minimum Wage" topic:

> B: "...you know welfare and minimum wage and and those type a things he cannot relate..."
> ...
> A: "...wage job i had a four year old that i put in daycare and after..."
> ...
> A: "...the difference on tips-"
> B: "dear i know i know and and sometimes-"
> A: "no minimum wage..."

3.4. PLSA Modeling with LexRank Extraction of 3 Best Frames

Latent topic models have proven to be effective in semi-supervised extractive summarization systems which utilize a classifier to extract salient sentences [13]. This work uses a different, unsupervised approach by incorporating a latent topic model into the LexRank algorithm. As in the first system, each node in the graph represents a frame, but now edge weights are computed using a cosine similarity measure between the latent topic score vectors defined in Eq. 4. The top 3 15-word frames are extracted using the MMR reranking scheme, resulting in a 45 word long summary. Below is an example summary generated by system 2 for the same conversation used in the example for system 1:

> B: "...along with their tips anyway"
> A: "yes"
> B: "i don't know being a waitress i've never been..."
> ...
> B: "...many aspects and you know welfare and minimum wage and and those type a things..."
> ...
> A: "...sad is the restaurants can get away with paying their waitresses two fifty an hour..."

3.5. PLSA Modeling with LexRank Extraction of Best Frame

The third system is a variation on the second, except rather than extracting the top 3 scoring frames to compose the summary, the single frame with the highest LexRank score is extracted. To maintain the same summary length as the previous two systems, 45 word long windows are used. The motivation for this system lies in addressing the question of whether three short snippets from different locations within a document are more informative than one longer, more coherent snippet. The summary generated by system 3 for the example conversation from the "Minimum Wage" topic is shown below:

> A: "...sad is the restaurants can get away with paying their waitresses two fifty an hour and expect them to make up the difference on tips"
> B: "dear i know i know and and sometimes-"
> A: "no minimum wage is nothing"
> B: "that that's true that's true and some..."

3.6. Latent Topic Modeling with Signature Word Extraction

The fourth system extracts and displays only the top 10 unique words that are most topically relevant to the document. The first step is to compute the topical LLR score vector for the entire document:

$$\vec{s}_d = \sum_{\vec{s}_i \in d} \vec{s}_i \tag{9}$$

and then for each word in the document:

$$\vec{s}_w = \left[\log \frac{P(w|z_1)}{P(w|\bar{z}_1)}, \ldots, \log \frac{P(w|z_{N_z})}{P(w|\bar{z}_{N_z})}\right]^T \tag{10}$$

The topical relevance of a word w in a document d can then be calculated using a count-weighted dot product similarity between the topical LLR score vectors for w and d:

$$\text{Rel}(w, d) = (c_{w,d})^{\frac{1}{2}} (\vec{s}_w \cdot \vec{s}_d) \tag{11}$$

Taking the square root of $c_{w,d}$ compresses the counts of all the words in d, thereby allowing rarer but still topically significant words to compete with very frequently occurring words in the document. The 10 words in d with the highest relevance scores are extracted to form a summary for the document. The summary generated by system 4 for the same "Minimum Wage" conversation is shown below:

> wage, minimum, welfare, daycare, tip, hour, fifteen, insurance, waitress, sufficient

4. EXPERIMENTS

4.1. Experimental Design for Amazon Mechanical Turk

Amazon's Mechanical Turk crowdsourcing service was used to perform an extrinsic evaluation of our summarization systems using human subjects. We created a collection of multiple-choice questions, each of which was comprised of a summary of one of the Fisher conversations using one of the 4 systems described in Section 3, followed by five topic prompts. Workers were tasked with reading the summary and then choosing the topic prompt which most likely started the conversation. For example, the prompt corresponding to the "Minimum Wage" topic was:

> *Do each of you feel the minimum wage increase - to $5.15 an hour - is sufficient?*

The prompts displayed for each question included the ground truth prompt, as well as the top 4 scoring incorrect prompts as determined by an automatic topic ID system [14] run on the full text of the source conversation. Each human intelligence task (HIT) contained 5 questions: one question for each of the 4 summarization systems, and a handcrafted verification question with an obvious answer. The source conversations used to create the summaries were guaranteed to be different within a HIT. To filter out low quality work, any HIT submitted with an incorrect answer to the verification question was rejected. 1374 unique HITs were posted to Mechanical Turk, and workers were paid $0.05 per HIT completed. Only workers from the United States who had a previous approval rating of 90% or higher were allowed to work on the task, and a total of 152 unique workers participated. We also included triple redundancy in our evaluation, i.e. each HIT was posted on Mechanical Turk until it was completed by 3 different workers who all passed the verification measures.

4.2. Experimental Results

In Table 1, we show the topic ID error rates for each summarization system used in the Mechanical Turk evaluation. In addition to the overall human error rates, we show the error rates corresponding to assigning a topic label based upon a majority vote between the 3 workers who evaluated each summary. In the rightmost column, the performance of an automatic topic ID system [14] is shown. This system was trained on a completely independent set of 1372 Fisher conversations, and then asked to identify the most likely prompted topic for each of the summaries seen by the humans. Using McNemar's test, we confirmed that the pairwise performance differences between summarization systems reflected in the machine and human majority vote columns are all statistically significant to a p-value of 0.05, with the single exception of system 2 vs. system 4 in the human majority vote column.

The relative performance rankings of the 4 systems are nearly identical according to the human and machine evaluations. The only discrepancy lies in the fact that system 4 achieved a lower error rate than system 2 in the machine evaluation, while these two systems effectively tied in the human study. This is not surprising, since

Table 1. Comparison of machine vs. human topic ID error rates on the summaries of 1374 Fisher conversations using the four systems discussed in Section 3, as well as the full text for the machine system

System	Avg. # words	Comp Ratio (%)	Human Overall ER (%)	Human Maj. Vote ER (%)	Machine ER (%)
1	45	2.47	18.7	15.9	19.5
2	45	2.47	15.0	13.3	14.6
3	45	2.47	21.0	18.6	25.4
4	10	0.55	16.1	13.8	11.6
Full Text	1825	100	N/A	N/A	4.3

the topic ID system used in these experiments worked under a bag-of-words assumption, and did not use any contextual information. Therefore, the machine system was naturally more suited towards the use of system 4 than the human subjects, who outperformed the machine system when context was included in the summaries. Otherwise, the agreement between the rankings implies that ID performance between humans and machines is correlated.

We can also make some key observations regarding the performance of the summarization systems relative to one another:

1. The selection of signature key words (system 4) not only produced the most condensed summarizations (with a summary size of only 10 words) it also yielded significantly better automatic topic ID accuracy than the other systems. While human topic ID accuracy using system 4 is effectively equivalent to system 2, the performance of system 4 is impressive considering its low compression ratio.

2. The PLSA based frame similarity measure used in system 2 outperformed the direct cosine similarity measure used in system 1. This indicates that the incorporation of a latent topic model into an extractive summarization system provides a considerable performance boost over a direct model.

3. The single frame extraction system using long frames (system 3) performed worse than extraction of individual utterances using either the topical frame similarity measure (system 2) or the direct similarity measure (system 1). This demonstrates that a larger collection of intelligently selected frames contains more topical information than a single frame of greater length and coherence.

5. CONCLUSIONS

In this paper we examined the use of an extrinsic evaluation paradigm for assessing and comparing different document summarization techniques. We explored the application of summarization systems to an information retrieval task, in which identifying the topic of a document based on a small extractive summary alone was considered to be equivalent to a relevancy detection task. We presented the results of a large-scale study performed using human subjects via Amazon's Mechanical Turk, and then showed that the performance of the human test subjects appeared to be correlated with the performance of an automatic topic identification system on the same task. This indicates that automatic topic identification systems can be utilized as low-cost, fast evaluation tools used during the research and development stages of summarization systems, especially those intended to be used in information retrieval systems.

In addition to this result, we have shown that incorporating a latent topic model into extractive summarization systems provides a considerable boost in performance over a system based on bag-of-word count vectors for a topic identification task. We have also shown that both human and machine topic ID performance based solely on a small set of signature words extracted without any surrounding context is very effective relative to summaries over 4 times greater in length generated using utterance extraction. This may be especially useful for speech-based summarization where counts of common keywords can be reliably estimated over an entire document, but extracted utterance snippets with errorful transcripts may be difficult for users to read and interpret.

6. ACKNOWLEDGEMENTS

The authors would like to thank Jim Glass and Jackie Lee of MIT CSAIL for their assistance in conducting the Mechanical Turk experiments.

7. REFERENCES

[1] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. of ACL 2004 Workshop: Text Summarization Branches Out*, 2004.

[2] K. McKeown, H. Jing, R. Barzilay, and M. Elhedad, "Summarization evaluation methods: Experiments and analysis," in *Proc. of AAAI Symposium on Intelligent Summarization*, 1998.

[3] D. House, L. Hirschman, T. Firmin, I. Mani, G. Klein and B. Sundheim, "SUMMAC: A text summarization evaluation," in *Natural Language Engineering*, vol. 8, pp. 43–68, 2002.

[4] C. Wayne, "Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation," in *Proc. of Int. Conf. on Language Resources and Evaluation*, 2000.

[5] C. Cieri, D. Miller and K. Walker, "The Fisher Corpus: A resource for the next generation of speech-to-text," in *Proc. of Int. Conf. on Language Resources and Evaluation*, 2004.

[6] C. Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk," in *Proc. of EMNLP*, 2009.

[7] D. Gillick and Y. Liu, "Non-expert evaluation of summarization systems is risky," in *Proc. of NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[8] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of 15th Conf. in Uncertainty in Artificial Intelligence*, 1999.

[9] T. Hazen, "Latent topic modeling for audio corpus summarization," in *Proc. of Interspeech*, 2011.

[10] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[11] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. of 21st Int. Conf. of the Association for Computational Linguistics*, 2006.

[12] G. Erkan and D. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457-479, 2004.

[13] A. Celikyilmaz and D. Hakkani-Tur, "Extractive summarization using a latent variable model," in *Proc. Interspeech*, 2010.

[14] T. Hazen, "MCE training techniques for topic identification of spoken audio documents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2451–2461, 2011.