The NTNU Summarization System at MultiLing 2015

Hisao-Tsung Hung, Kai-Wun Shih and Berlin Chen

Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
{80347004S, 60247065s, berlin}@ntnu.edu.tw

Abstract

This paper describes the empirical results obtained by leveraging an unsupervised statistical representation and modeling framework for the Call Center Conversation Summarization (CCCS) task at Multiling'15. In particular, the characteristics and performance levels of various summarization methods originating from such a framework are analyzed and compared. Nevertheless, the results of our first participation in the CCCS evaluation seem to have room for improvement, which awaits further studies.

1 Introduction

In the recent past, we have witnessed a flurry of research activity aimed at the development of novel and effective methods for speech summarization (Liu and Hakkani-Tur, 2011; Liu et al., 2015), which purports to generate a concise summary that can help users efficiently review and quickly digest the important information conveyed by either a single spoken document or multiple spoken documents. This is attributed in large part to the great progress in automatic speech recognition (ASR) and overwhelming growth of multimedia associated with spoken documents, such as broadcast news and recordings of conversations, among others, made available to the public (Lee and Chen, 2005; Furui et al., 2012).

Despite the preliminary success on some application tasks, it is believed that speech summarization is still far from being solved. For example, speech summarization inevitably suffers from the problems of recognition errors and incorrect sentence boundaries when using ASR techniques to transcribe the spoken documents into text forms. On the other hand, speech summarization also presents information cues that are peculiar to it and do not exist for text summarization, such as information cues about prosodies/acoustics and emotions/ speakers, which can potentially help in determining the important parts or implicit structures of spoken documents.

The generation process of a summary basically can be either abstractive or extractive. In abstractive summarization, a fluent and concise abstract that reflects the key concepts of a document is generated, whereas in extractive summarization, the summary is usually formed by selecting salient sentences from the original document. The former requires more sophisticated natural language processing (NLP) techniques, including semantic representation and inference, as well as natural language generation, while this would make abstractive approaches difficult to replicate or extend from constrained domains to more general domains. Apart from being abstractive or extractive, a summary may also be generated by considering several other aspects like being generic or query-oriented summarization, single-document or multi-document summarization, among others (Mani and Maybury, 1999; Liu and Hakkani-Tur, 2011; Nengova and McKeown, 2011).

In this paper, we present an unsupervised statistical representation and modeling framework for use in summarizing call center conversations, which involves two major processing steps. In the first step, a spoken document (or conservation recording) to be summarized and each of its constituent sentences are appropriately (or concisely) represented in vector form based on some statistical features of word usage and/or co-occurrence relationship. As such, important sentences can be selected as the candidates to be included in the summary on the basis of a relevance measure between the document and each of its sentences. In the second step, we exploit a greedy sentence compression mechanism to reduce the number of superfluous words or text segments among the selected sentences.

The remainder of this paper is structured as follows. We start by an elucidation of the proposed modeling framework and associated methods in Section 2. After that, the experimental setup and a series of experiments and associated

discussions are presented in Sections 3. Finally, Section 4 concludes our presentation and discusses avenues for future work.

2 Proposed summarization framework

We propose an unsupervised statistical modeling framework for use in summarizing call center conversations, which involves two important processing steps: 1) sentence representation and selection; 2) sentence merging and compression. In the following, we shed light on the details about these two processing steps.

2.1 Sentence representation and selection

We apply the vector space method (VSM), well-practiced in the information retrieval (IR) community, to the speech summarization task studied in this paper (Baeza-Yates and Ribeiro-Neto, 2011). VSM represents each sentence of a document, and the whole document, in vector form. In this method, each dimension specifies the weighted statistics, for example the product of the term frequency (TF) and inverse document frequency (IDF) (Baeza-Yates and Ribeiro-Neto, 2011), associated with a word in the sentence or the document. Sentences are ranked in descending order on the basis of the cosine relevance score calculated between the vector of each sentence and the vector of the whole document. By doing so, sentences with the highest relevance scores to the whole document are regarded as the candidates to be included in the summary.

Apart from VSM, we alternatively investigate two word embedding methods, i.e., the continuous bag-of-word (CBOW) method (Mikolov et al., 2014) and the paragraph vector (PV) method (Le and Mikolov, 2014), for use in speech summarization, which have recently shown excellent performance in many natural language processing (NLP) related tasks, such as sentiment analysis and sentence completion. The common notion of these methods is to learn fixed-length continuously distributed vector representations of words (or sentences or documents) using neural networks, which aim to probe latent semantic and/or syntactic cues that can in turn be used to induce similarity measures among words, sentences and documents. The structure of CBOW bears a close resemblance to a feed-forward neural network, with the exception that the non-linear hidden layer in the former is removed. By doing so, the model can still retain good performance and be trained on much more data efficiently while getting around the heavy computational burden incurred by the non-linear hidden layer. Formally, given a sequence of words, $w^1,w^2,\ldots,w^T$, the objective function of CBOW is to maximize the log-probability expressed as follows:

$$\sum_{t=1}^{T} \log P(w^t \mid w^{t-c},...,w^{t-1},w^{t+1},...,w^{t+c}), \quad (1)$$

where c is the window size of the training context for the central word w^t; T denotes the length of the training corpus. The conditional probability in Eq. (1) is defined by

$$P(w^t \mid w^{t-c},...,w^{t-1},w^{t+1},...,w^{t+c}) = \frac{\exp(\mathbf{v}_{\overline{w}^t} \cdot \mathbf{v}_{w^t})}{\sum_{i=1}^{V} \exp(\mathbf{v}_{\overline{w}^t} \cdot \mathbf{v}_{w_i})}, \quad (2)$$

where $\mathbf{v}_{w^t}$ denotes the vector representation of the word w at position t; V indicates the size of the vocabulary; and $\mathbf{v}_{\overline{w}^t}$ denotes the (weighted) average of the vector representations of the context words of w^t, which can be further expressed in the form:

$$\mathbf{v}_{\overline{w}^t} = \sum_{j=-c, j\neq 0}^{c} \alpha_j \mathbf{v}_{w^{t+j}}, \quad (3)$$

where α_j is a weighting factor associated with the distance between the central word w^t and the context word w^{t+j}. The concept of CBOW is motivated by the distributional hypothesis (Millera and Charles, 1991), which states that words with similar meanings often occur in similar contexts and thus suggests to look for word representations that capture their context distributions.

Once the vector representation of each individual word is estimated, the vector representation of a sentence (or a document) by averaging the vector representations of words occurring in the sentence (or the document). Along the same vein as VSM, sentences can be ranked in accordance with the cosine relevance score calculated between the CBOW-based vector representation of each sentence and that of the whole document.

On the other hand, the PV method attempts to learn the fixed-length distributed vector representations of pieces of texts in a direct manner, where the texts can be of variable-length sentences, documents, and more. As an illustration, the vector representation of a sentence is estimated by maximizing the likelihood of predicting all words involved in the sentence; that is, the sentence vector is concatenated with several word vectors from the sentence and predicting the following word in a given context (Le and Mikolov, 2014). The document vector of the

document to be summarized can be estimated in a similar manner. Following, sentences can be ranked in accordance with the cosine relevance score calculated between the PV-based vector representation of each sentence and that of the whole document.

2.2 Sentence compression

After the construction of a ranked list of candidate summary sentences, we subsequently employ a greedy sentence compression mechanism that attempts to recursively select and compress any pair of candidate sentences (from the top of the list) exhibiting redundancy in word usage that is greater than a predefined threshold through a word graph construction and decoding process.

3 Experiments

3.1 Experimental setup

The summarization dataset used in the CCCS task at Multiling'15 consists of two subsets of spoken documents respectively drawn from French and an Italian call center conversation recordings, i.e., the Decoda corpus and Luna corpus. For the Decoda corpus, a subset of 1000 French call center conversations without corresponding human-generated abstractive summaries and 50 conversations with corresponding human-generated abstractive summaries was provided for training the associated models of the various summarization methods. Another subset of 100 conversations with corresponding human-generated abstractive summaries was reserved as the test set.

On the other hand, for the Luna corpus, a subset of 261 Italian conversations without corresponding human-generated abstractive summaries and 100 conversations with corresponding human-generated abstractive summaries was provided for training the associated models of the various summarization methods. Another subset of 100 conversations with corresponding human-generated abstractive summaries was reserved as the test set. In addition, for both the Decoda corpus and the Luna corpus, their test sets were also translated to English for further performance evaluation (denoted by the EN corpus). We refer readers to (Favre et al., 2015) for an elaborate description of the CCCS task.

The summarization performance of the various summarization methods compared in this paper is evaluated with widely-used ROUGE-2 metric (Lin, 2004), which computes the recall in terms of word bigram overlaps between a set of reference (human-generated) abstractive summaries and the automatic summaries output by each summarization method.

	Decoda	Luna	EN
VSM	0.035	0.013	0.023
CBOW	0.027	0.015	0.031
PV	0.034	0.012	0.024
Baseline-MMR	0.045	0.020	0.029
Baseline-L	0.040	0.015	0.023
Baseline-LB	0.046	0.027	0.025

Table 1: The results achieved by our three proposed summarization methods and three baseline methods.

3.2 Experimental results and discussion

The summarization results obtained by our presented methods, i.e., VSM, CBOW and PV are shown in Table 1, where the summarization results of three baseline summarization methods were also listed for performance comparison. The first baseline method is the maximal marginal relevance (Carbonell and Goldstein, 1998) method (denoted by Baseline-MMR); the second one simply selects the first words of the longest turn in the conversation, up to the length limit, as the summary (denoted by Baseline-L); and the third one selects the word of the longest turn in the first 25% of the conversation corresponding to the description of the caller's problem (denoted by Baseline-LB). Inspection of Table 1 reveals two noteworthy points. On one hand, among the three methods that we have developed in this paper, CBOW stands out in performance for the Luna and EN corpora; however, the situation is reversed for the Decoda corpus. The reason behind such a discrepancy in performance is to be examined thoroughly in our future work. On the other hand, our presented summarization methods cannot match the performance levels of the three baseline methods, except for the EN corpus where CBOW beats the three baseline methods by a significant margin. Again, it still awaits further study to enhance and extend our methods for the CCCS task or other more complicated speech summarization tasks.

4 Conclusions and future work

In this paper, we have reported on the empirical results obtained by leveraging an unsupervised statistical representation and modeling frame-

work for the Call Center Conversation Summarization (CCCS) task at Multiling'15. In addition, the characteristics and performance levels of various summarization methods stemming from such a framework have been are analyzed and compared. However, the results of our first participation in the CCCS evaluation are far from being satisfied compared to the baseline methods; the reasons behind this await further studies. We also list below two possible future extensions: 1) exploring more effective sentence selection and compression strategies, 2) incorporating more acoustic/prosodic, lexical and structural cues into the speech summarization methods (Chen et al., 2013).

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval: The Concepts and Technology behind Search*, ACM Press.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity based reranking for reordering documents and producing summaries. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336.

Berlin Chen, Shih-Hsiang Lin, Yu-Mei Chang, Jia-Wen Liu. 2013. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing & Management*, 49(1): 1–12, 2013.

Benoit Favre, Evgeny Stepanov, J'er'emy Trione, Fr'ed'eric B'echet and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. Submitted to *the 16th Annual SIGdial Meeting on Discourse and Dialogue*.

Sadaoki Furui, Li Deng, Mark Gales, Hermann Ney, and Keiichi Tokuda. 2012. Fundamental technologies in modern speech recognition. *IEEE Signal Processing Magazine*, 29(6):16–17.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In P*roc. of International Conference on Machine Learning*.

Lin-Shan Lee and Berlin Chen. 2005. Spoken document understanding and organization. *IEEE Signal Processing Magazine*, 22(5):42–60.

Chin-Yew Lin. 2004. ROUGE: a Package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out*.

Shih-Hung Liu, Kuan-Yu Chen, Berlin Chen, Hsin-Min Wang, Hsu-Chun Yen, Wen-Lian Hsu. 2015. Combining relevance language modeling and clarity measure for extractive speech summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6): 957–969.

Yang Liu and Dilek Hakkani-Tur. 2011. *S*peech Summarization. Chapter 13 in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, G. Tur and R. D. Mori (Eds), Wiley, New York.

Inderjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In P*roc. of International Conference on Learning Representations*: 1–12.

George A. Millera and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2 – 3):103–233.