Spotting keywords and sensing topic changes in speech

Zhu, Xiaodan

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/CISDA.2012.6291537
CISDA 2012: Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defence Applications, pp. 1-7, 2012-07-13

Spotting keywords and censoring topic changes in speech

Xiaodan Zhu

National Research Council of Canada, Ottawa, Canada
Xiaodan.Zhu@nrc-cnrc.gc.ca

Abstract

Security concerns involved in dealing with sensitive information conveyed in human languages cannot circumvent speech, which is the most basic, natural form of human communication and a huge amount of data are generated daily. Dealing with such data is naturally associated with typical big-data problems in terms of both computational complexity and storage space. Unfortunately, compared with written texts, speech is inherently more difficult to browse, if no technical support is provided. In this paper we are interested in spotting keywords, which could reflect a security agent's information needs, and study its usefulness in helping automatically disclose topic changes (boundaries) in speech data under concern. Our results show that keyword spotting can help identify topics with a competitive performance.
Index Terms: keyword spotting, topic segmentation, speech understanding.

1. Introduction

Security applications concerned with finding sensitive information conveyed in human languages should not circumvent speech, which is the most basic, natural form of human communication. Information exchanged between two or more persons, particularly that with an explicit secret nature, is very likely through voice media, compared with being put down in a written form instead. In addition, the consistently increasing availability of spoken content in social media and other digital sphere provides additional, considerable opportunities to mine security-related information. Unfortunately, speech is inherently much more difficult to browse than texts, due to its more linear or sequential property in the traditional delivery, if no additional technological support is provided. In general, dealing with spoken content is inherently associated with problems typical to big data in terms of computational complexity and storage space involved.

Keyword spotting is a basic technology to help people find the information they are interested in, in which speech forms of keywords, often expressed in a frequency space, are compared with the corresponding speech archives to find their occurrences. On the other hand, knowing keywords' location in speech data does not guarantee an accurate understanding of content of speech, where words are connected to form semantic or discourse cohesion and coherence in expressing things that one may be interested in; keyword spotting could, however, help further disclose more structures and information.

The very basic yet most intensively studied approach to represent a document is through topic segments, where a document is linearly segmented to topically or subtopically cohesive segments. Each segment of interest can then be accessed by human or other automatic applications such as information retrieval. Though topic segmentation is conceptually very simple, it is important, particularly for spoken documents. Compared with written texts, which are almost always presented as more than uninterrupted strings of texts (e.g., with manually created paragraph boundaries, sections/subsections, chapters, and their titles), the general lack of semantic formalities in speech data, accompanied by the inherent difficulty of browsing, results in more prominent usefulness of inferring any forms of semantic or discourse structures, compared with the situation in its written counterparts. This would include the very basic topic/subtopic segmentation as well as hierarchical topic segmentation, given the often subtle and fine-grained topic or subtopic shifting in spoken archives.

In this paper, we provide experimental evidence in studying the interaction between these two basic problems; that is, we attempt to understanding the usefulness of keyword spotting in helping disclose topic changes in the speech data under concern. Our results show that keyword spotting can help identify topics with a competitive performance.

2. Related work

Topic segmentation has received its most intensive study on written texts. Unlike in the typical case of a multi-paragraph phenomenon happening [1] above paragraphs but under low-level semantic markings, topic segmentation in spoken documents is entangled more closely with semantic hierarchies, and is usually more subtle with the often fine-grained topic shifting. Method-wide, a variety of models and features have been proposed, among which lexicon-cohesion based models are possibly the most prominent ones: lexical cohesion underlies

many other feature-based models and by itself can often achieves the-state-of-the-art performance, particularly for fine-grained topics [2]. Such models are generally more independent of specific styles, genres, and speakers of documents. Instead of considering cohesion in speech itself, in this paper, we consider the security agents' information needs: each topic/subtopic they are interested in are expressed with a small set of keywords (e.g., *Yemen explosion in May* and *the consequence of the explosion*). Content words in these two descriptions are used in keyword spotting to guild the process of finding these two subtopics in a speech document that contains them. We will discuss more details of the model later.

Because of its performance and confound connection with other models, we adopt the cohesion-based approach to understand our problems. The specific model we leverage [3] considers not just local but also long-range cohesion dependencies, based on a graph partitioning framework, which improves segmentation accuracy and is robust to speech recognition errors. More specifically, we will introduce the extension of the model [4, 5] in Section 3, which can incorporate the keywords to reflect security agents' information needs.

In a even more general setting, analyzing discourse structures can provide thematic skeletons (often represented as trees) of a document as well as relationship between the nodes in the trees. Examples include the widely known discourse parsing work of [6], among others. However, when the task involves the understanding of high-level discourse, it becomes more challenging than finding local discourse conveyed by small spans of texts; e.g., the latter is more likely to benefit from the presence of discourse markers. Specifically for spoken documents, speech recognition errors, absence of formality and thematic boundaries, and less linguistically well-formedness of the spoken language, will further impair the conditions on which a reliable discourse analysis algorithm is often built. In stead of addressing the ultimate goal of automatically inferring hierarchical structures for spoken documents, this paper focuses more on further understanding linear segmentation.

3. A integrated graph-partitioning framework

As discussed above, the work of [3] proposes a state-of-the-art topic segmentation model for spoken documents based on global lexical cohesion. In a more general setting, the work of [4, 5] proposes a graph-partitioning-based framework to solve a semantic tree-to-string alignment problem, which subsumes the topic segmentation as a submodel and incorporates an additional alignment submodel; all the formulation is kept in a graph-partitioning framework. We leverage this framework to understand our problems here: we utilize keyword spotting to establish similarities between the designated topic description and the corresponding utterances in a speech document; we then leverage the alignment to find topic shifting/segmentation boundaries corresponding to these topics under concern.

As in a simple example shown in Figure 1, we have a sequence of speech utterances, $u_1, ..., u_8$, with similarities associating among them, denoted by the green dotted lines. This part of sub-graph, i.e., utterances and their similarities, forms a square similarity matrix, which is used by [3] to build the topic segmentation model. Our focus here, however, is the upper subgraph in the figure, i.e., that formed by the utterances , the topic description items $L_1 ... L_3$, and their similarities (dotted red lines) forms a bipartite graph. We utilize these part to incorporate keywords; i.e., each node L_i is a short description about a topic that an agent or an application is interested in, and the description is composed of several keywords, the occurrences of these keywords in the speech can then be found by using keyword spotting method; with which the similarities between the topic descriptions and the utterances in speech can be established, which can then be used to choose the proper levels/details topic segments corresponding to the description. We leverage graph-partitioning models to address this problem. Note that, the advantage of using keyword spotting lie in that this process does not need transcripts of speech. Our results show that keyword spotting can help identify topics with a competitive performance comparable to those needing full transcription of speech. Note also that in Figure 1, the similarities between utterances and topic descriptions are shown as binary, while in real model computation, all similarity scores are real values, calculated as we discuss later.

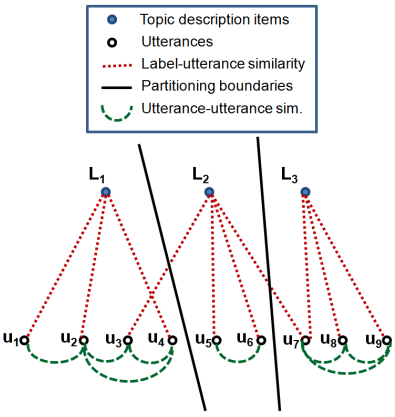

Figure 1: An example of the graph-partitioning framework.

The model is formulated in a unified graph-partitioning framework. Consider a general, simple two-set partitioning case, in which a boundary is placed on a graph $G = (V, E)$ to separate its vertices V into two sets,

A and B, with all the edges between these two sets being removed. The objective, as we have mentioned above, is to minimize the following normalized-cut score:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \quad (1)$$

where,

$$cut(A,B) = \sum_{a\in A, b\in B} w(a,b)$$
$$assoc(A,V) = \sum_{a\in A, v\in V} w(a,v)$$
$$assoc(B,V) = \sum_{b\in B, v\in V} w(b,v)$$

In equation (1), $cut(A,B)$ is the total weight of the edges being cut, i.e., those connecting A with B, while $assoc(A,V)$ and $assoc(B,V)$ are the total weights of the edges that connect A with all vertices V, and B with V, respectively. In general, minimizing such a normalized-cut score has been shown to be NP-complete. In the specific setting here, however, the solution is constrained by the linearity of segmentation on transcripts, a polynomial-time algorithm exists.

Back to Figure 1 above, we focus on the bipartite graph in the upper part. To find utterance spans corresponding to the topic-description items, we place $m-1$ boundaries onto the bipartite graph to partition the graph into m bipartite graphs and obtain triples, e.g., (L_i, u_j, u_k), to align L_i to $u_j, ..., u_k$, where $L_i \in \{b_1, ..., b_m\}$ and $u_j, u_k \in \{u_1, ..., u_n\}$ and $j <= k$. The optimal solution maximizing the normalized minimum cut score discussed above can be found with a dynamic-programming process with a recurrence relation:

$$C[i,k] = \min_{j\le k}\{C[i-1,j] + D[i,j+1,k]\} \quad (2)$$

In equation (2), $C[i,k]$ is the optimal/minimal normalized-cut value of aligning the first i topic labels, $L_1, ..., L_i$, with the first k utterances, $u_1, ..., u_k$. It is computed by updating $C[i-1,j]$ with $D[i,j+1,k]$, for all possible j s.t. $j \le k$, where $D[i,j+1,k]$ is a normalized-cut score for the triple (b_i, u_{j+1}, u_k) and is defined as follows:

$$D[i,j+1,k] = \frac{cut(A_{i,j+1,k}, V \setminus A_{i,j+1,k})}{assoc(A_{i,j+1,k}, V)} \quad (3)$$

where $A_{i,j+1,k}$ is the vertex set that contains the bullet b_i (including its descendant bullets, if any, as discussed above) and the utterances $u_{j+1}, ..., u_k$; $V \setminus A_{i,j+1,k}$ is its complement set.

$$C[i,k] = \min_{j\le k}\{C[i-1,j] + \lambda_1 D[i,j+1,k] + (1-\lambda_1)S[j+1,k]\} \quad (4)$$

where,

$$S[j+1,k] = \frac{cut(A_{j+1,k}, V \setminus A_{j+1,k})}{assoc(A_{j+1,k}, V)} \quad (5)$$

4. Experiment set-up

Corpus We chose to use presentation recordings as the experimental data to address our problem here, since for each presentation, we can utilize those words on bullets of the electronic slides to simulate topics that attract security agents: each bullet is regarded as a short topic description that a security agent is interested in to know: what is discussed in the speech about these topics? As discussed above, words in each bullet are regarded as keywords and are used to establish similarity between each bullet and utterance in speech, with keyword spotting method discussed below.

In addition to this convenience, presentation is colloquial and full of spoken-language characteristics such as disfluencies, and is therefore an ideal data to simulate other domains that a security agency is really interested in, e.g., telephone conversations, speech in social media, meeting recordings, etc. Note that news broadcasts that are often used to develop and test speech recognition systems are less ideal in our problems here for these reasons.

Specifically, our experiment uses a corpus of four 50-minute university lectures taught by the same instructor, which contain 119 slides composed of 921 bullets. The automatic transcripts of the speech contain approximately 30,000 word tokens, roughly equal to a 120-page double-spaced essay in length. The lecturer's voice was recorded with a head-mounted microphone with a 16kHz sampling rate and 16-bit samples, while students' comments and questions were not recorded. The speech is split into utterances by pauses longer than 200ms, resulting in around 4000 utterances. Each lecture is divided into three parts if roughly the same length to speed up computation, so we have 12 lecture parts in total.

Keyword spotting In this paper, we use a token-passing based algorithm provided in the ASR (automatic speech recognition) toolkit SONIC [7]. Since the slides are given in advance, we manually add into the pronunciation dictionary the words that appear in slides but not in the dictionary. To estimate similarity between a word vector and an utterance, we sum up all keyword-spotting confidence scores assigned between them, normalize the resulted score by the length of the vector and the duration of the utterance, and then renormalize it to the range [0, 1] within the same spoken lecture.

To understand the effective of keyword spotting in finding topic boundaries, we compare the results with those achieved on automatic transcripts. The transcripts were generated with the SONIC toolkit [7], with the models trained as suggested by [8], in which one language model was trained on SWITCHBOARD and the

other used also corpus obtained from the Web through searching the words on slides, which result in a 48% and 43% word error rate (WER), respectively. Both bullets and automatic transcripts were stemmed with the Porter stemmer and stopwords were removed. The similarities between bullets and utterances and those between utterances were calculated with different distance metrics, i.e., cosine, exponential cosine [3] for topic segmentation, and a normalized word-overlapping score used in summarization [9], from which we chose the one (regular cosine) that optimizes our baseline. Our graph-partitioning models then used exactly the same setting. The lexical weighting is same as in [3], for which we split each lecture into M chunks, the number of bullets. Finally, we obtained a M-by-N bullet-utterance similarity matrix and a N-by-N utterance-utterance matrix to optimize the alignment model and topic-segmentation model, respectively, while M and N, as already mentioned, denote the number of bullets and utterances of a lecture, respectively.

Evaluation metric The metric used in our evaluation is straightforward—automatically acquired boundaries on transcripts for each slide bullet are compared against the corresponding gold-standard boundaries to calculate offsets measured in number of words, so the smaller the value is, the better the performance is. Note that topic segmentation research often uses metrics such as P_k and WindowDiff [3, 10, 11]. Our problem here, however, has an exact 1-to-1 correspondence between a gold and automatic boundary, in which we can directly measure the exact word offset of each boundary.

5. Results

We evaluate the system performance in a comprehensive scenario, in which, instead of using a flat level of topic descriptions, we use all topic descriptions (all bullets on lecture slides) on multiple levels organized in a hierarchical structure; by nature, a larger topic can include several subtopics, and we leave the flexibily of deciding the granularity of topics to the security agents or other applications, specified by his/her specific topic descriptions that could be flat (or not). In other words, for a comprehensive understanding, we evaluate all levels of topics specified by the bullet hierarchies in a lecture all together, of which topic segmentation is just a special case (with flat topic structure). Specifically, we compare the effectiveness of keyword spotting with the performance obtained by using full transcripts in six typical models: *SeqBase*, *SeqCut*, *HieBase*,*HieCut*,*PrsBase*, and *PrsCut*. Details about the six models can be found in [5]. Here, the readers may just assume them as black boxes: three (*SeqCut*, *HieCut*, and *PrsCut*) are modeling variants that utilize the graph-partitioning framework discussed above, and the other three are models that utilize traditional, well-known dynamic time warping (*SeqBase*, *HieBase*, and *PrsBase*). Again, the smaller a score is, the better the performance is, since the a score represents offset errors.

In Table 1, the *KWD* column represents the experimental results obtained with the token-pass keyword spotting technology as discussed above, without any transcription of speech. To show its effectiveness, we compare the results with those achieved on full transcripts with a WER of 0.43 and 0.48, which are typical for lectures and conference presentations in realistic and less controlled situations [8, 12]. As a reference, we also list the results achieved on manual transcripts (the *Man* column). The table shows the competitive performance of using keyword spotting in this task: its performance is better than that observed on transcripts with a 0.48 word error rate. More exactly, for all these six models, the topic-boundary performance of using keyword spotting is narrowed down in between those observed on transcripts with a WER of 0.43 and 0.48.

	Lectures			
	Man.	WER=.43	KWD	WER=.48
(1) SeqBase	13.64	15.19	**16.18**	18.44
(2) SeqCut	10.35	12.87	**15.19**	16.16
(3) HieBase	19.72	21.06	**23.73**	24.25
(4) HieCut	9.67	12.13	**15.71**	15.95
(5) PrsBase	13.43	15.05	**17.23**	18.18
(6) PrsCut	9.49	12.05	**14.18**	15.20

Table 1: The performances of different segmentation models.

6. Conclusions

As speech being the most basic, natural form of human communication and as a huge amount of spoken data being generated daily, security concerns involved in dealing with sensitive information conveyed in human languages should not circumvent speech. In this paper we are interested in two basic problems in help access spoken documents: one helps find information in concern (keywords spotting) and the other helps present spoken document better (topic segmentation). We leverage the former, keywords spotting, to incorporate security agents' or other applications' information needs in help censoring topic changes in speech. For this purpose, we employ a state-of-the-art token-pass keyword-spotting method to identify the occurrences of the keywords in the corresponding speech archives, with which we utilize a graph-partitioning approach to find the corresponding topic/subtopic boundaries through optimizing a normalized-cut criterion. Our experimental results show that the keyword spotting technology, without using

any transcripts, can achieve a competitive performance in finding topic changes/boundaries, comparable to those achieved on full transcripts with a typical range of speech recognition errors.

7. References

[1] M. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[2] P. Hsueh, J. Moore, and S. Renals, "Automatic segmentation of multiparty dialogue," in *Proc. of EACL*, 2006.

[3] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. of ACL*, 2006.

[4] X. Zhu, "A normalized-cut alignment model for mapping hierarchical semantic structures onto spoken documents," in *Proc. of CONLL*, 2011.

[5] X. Zhu, C. Cherry, and G. Penn, "Indexing spoken documents with hierarchical semantic structures: Semantic tree-to-string alignment models," in *Proc. of IJCLP*, 2011.

[6] D. Marcu, "The theory and practice of discourse parsing and summarization." The MIT Press, 2000.

[7] B. L. Pellom, "Sonic: The university of colorado continuous speech recognizer," *Tech. Rep. TR-CSLR-2001-01, University of Colorado*, 2001.

[8] C. Munteanu, G. Penn, and R. Baecker, "Web-based language modelling for automatic lecture transcription," in *Proc. of Interspeech*, 2007.

[9] D. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, pp. 919–938, 2004.

[10] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, no. 1-3, pp. 177–210, 1999.

[11] L. Pevsner and M. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Computational Linguistics*, vol. 28, pp. 19–36, 2002.

[12] B. Hsu and J. Glass, "Style and topic language model adaptation using hmm-lda," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2006.