ISSN: 2321-7782 (Online)
Impact Factor: 6.047
Volume 4, Issue 5, May 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Keyword Extraction for Document Recommendation in Conversation

Divya R K[1]
M tech Student
Dept.of Computer Science and Engineering
M Dason Institute of Engineering,Ulliyeri, Calicut
Kerala – India

Neethu Asokan[2]
Assistant Professor
Dept.of Computer Science and Engineering
M Dason Institute of Engineering,Ulliyeri, Calicut
Kerala – India

Vinitha V[3]
Assistant Professor
Dept.Computer Science and Engineering
M Dason Institute of Engineering,Ulliyeri, Calicut.
Kerala – India

Abstract: Document recommendation from ASR is a terribly troublesome task. This paper addresses the issue of keyword extraction from conversations. Using these keywords to retrieve, one significant potentially relevant document, this can be recommended to participants. It is difficult to infer precisely the information needs of the conversation participants. In this paper we specify how document recommended from conversation through ASR output. We first propose a diverse keyword extraction technique to extract keywords from the output of an ASR system. Then, we propose a method to derive multiple topically separated queries from this keyword set, in order to maximize the chances of making at least one relevant recommendation when using these queries to search over English Wikipedia.

Keywords: ASR, Keyword Extraction, Ranking, Document Retrieval, Document Recommendation.

I. Introduction

People are surrounded by an unprecedented wealth of information, available as documents, databases, or multimedia resources. Access to this information is conditioned by the availability of suitable search engines, but even when these are available, users often do not initiate a search, because their current activity does not allow them to do so, or because theyare not aware that relevant information is available [1].

The goal is to maintain multiple hypotheses about users information needs, and to present a recommendation based on the most likely ones. The recommending documents are related to users' current activities. Suppose when users participate in a meeting, their information needs can be modelled as implicit queries that are constructed from the pronounced words, obtained through real-time Automatic Speech Recognition (ASR) [6].

Therefore, aim at extracting a relevant and diverse set of keywords, cluster them into topic-specific queries ranked by importance, and present participants the most relevant result from these queries. The topic-based clustering decreases the chances of including ASR errors into the queries, and the diversity of keywords increases the chances that at least one of the recommended documents answers a need for information, or can lead to a useful document when following its hyperlinks.

In this paper, we introduce a novel diverse keyword extraction technique [3] from ASR output, which maximizes the coverage of potential information needs of users and reduces the number of irrelevant words. Once a set of keywords is extracted, it is clustered in order to build several topically-separated queries, which are run independently, by using the ranking keywords. The highest rank set is used to retrieve the recommended document.

The following sections, In Section II deals with related work. Sections III introduce proposed method. Section IV relates the result and discussions. Conclusion is summarized in section V. Section VI contains the papers, books, referred during the preparation of this paper.

II. RELATED WORK

Just-in-time retrieval systems have the potential to bring a radical change in the process of query-based information retrieval. Such systems continuously monitor users' activities to detect information needs, and pro-actively retrieve relevant information. To achieve this, the systems generally extract implicit queries (not shown to users) from the words that are written or spoken by users during their activities. In this section, review existing just-in-time-retrieval systems and methods used by them for query formulation. Also discuss previous keyword extraction techniques from a transcript.

Human's interactions with everyday productivity applications (e.g. word processors, Web browsers, etc.) provide rich contextual information that can be leveraged to support just-in-time access to task-relevant information. As evidence for claim, here present Watson [11], a system which gathers contextual information in the form of the text of the document the user is manipulating in order to proactively retrieve documents from distributed information repositories. This system close by describing the results of several experiments with Watson, which shows it consistently, provides useful information to its users.

The AMIDA Automatic Content Linking Device (ACLD) [10] is a just- in-time document retrieval system that constantly retrieves items from a repository and displays them to a participant or to all of them. The repository includes meeting related documents together with excerpts from previous meetings of the group. The device can be used online during a meeting, but also offline, integrated in a meeting browser.

A Speech-based Just-in-Time Retrieval System monitors an ongoing conversation or a monologue and enriches it with potentially related documents, including multimedia ones, from local repositories or from the Internet. The documents are found using keyword-based search or using a semantic similarity measure [9] between documents and the words obtained from automatic speech recognition. Results are displayed in real time to meeting participants, or to users watching a recorded lecture or conversation.

Several methods have been proposed to automatically extract keywords from a text, and are applicable also to transcribed conversations. The earliest techniques have used a new keyword extraction algorithm that applies to a single document [8] without using a corpus. Frequent terms are extracted first, and then a set of co-occurrence between each term and the frequent terms, i.e., occurrences in the same sentences, is generated. Co-occurrence distribution shows importance of a term in the document as follows. If probability distribution of co-occurrence between the term a and the frequent terms is biased to a particular subset of frequent terms, then the term a is likely to be a keyword.

Manual assignment of high quality keywords is expensive, time-consuming, and error prone. Therefore, most algorithms and systems aimed to help people perform automatic keywords extraction [18] have been proposed. Conditional Random Fields (CRF) [7] model is a state-of-the-art sequence labelling method, which can use the features of documents more sufficiently and effectively. At the same time, keywords extraction can be considered as the string labelling. Many unsupervised approaches [5]

The AMI(DA) system is a meeting room speech recognition system that has been developed and evaluated in the context of the NIST Rich Text (RT) evaluations. Recently, the Distant Access requirements of the AMIDA project have necessitated that the system operate in real-time. Another more difficult requirement is that the system fit into a live meeting transcription scenario. The AMI system for meeting room recognition is a combination of beam-forming, diarisation and ASR in real time [6].

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 [4], measured by the ability of the summary to aid a human in the completion of a specific task. This uses 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.

An improved method for keyword extraction from conversations, rewards both word similarity-to extract the most representative words, and word diversity-to cover several topics if necessary[3] were introduced. Inspired from summarization, the method maximizes the coverage of topics, those are recognized automatically in transcripts of conversation fragments.But it was unale to retrieve documents.

The present system addresses the problem of building concise, diverse and relevant lists of documents [2], which can be recommended to the participants of a conversation to fulfil their information needs without distracting them. These lists are retrieved periodically by submitting multiple implicit queries derived from the pronounced words. Each query is related to one of the topics identified in the conversation fragment preceding the recommendation, and is submitted to a search engine over the English Wikipedia. Here developed an algorithm for diverse merging of these lists, using a submodular reward function that rewards the topical similarity of documents to the conversation words as well as their diversity.

III. Proposed System

A. Proposed Outline

The overview of the proposed system is shown in Fig.1. With the use of ASR we tend to acknowledge the voice/speech from oral communication and convert it into text format. Then apply keyword extraction strategies to extract keyword from the text, after that clustering and ranking method are done for getting highest ranked keyword. With the premise of keywords we tend to use Just-in-Time Retrieval Systems [10][11][13] to retrieve relevant data. After retrieval we recommend data or information to users.

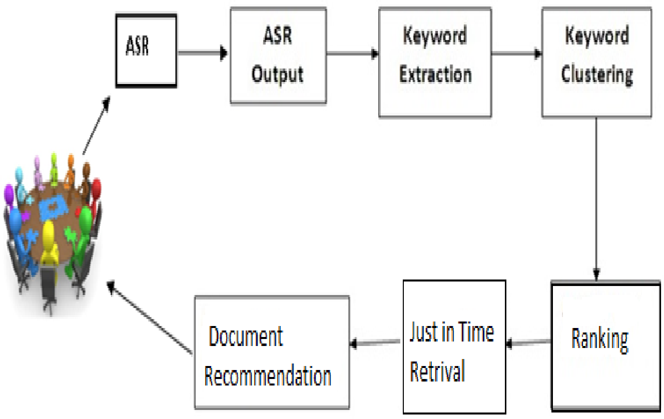

Fig. 1 Overview of Proposed System

We propose a two-stage approach to the formulation of implicit queries. The first stage is the extraction of keywords from the transcript of a conversation fragment for which documents must be recommended, as provided by an ASR system. These keywords should cover as much as possible the topics detected in the conversation, and if possible avoid words that are obviously ASR mistakes. The second stage is the clustering of the keyword set in the form of several topically-disjoint queries .Before going for the first stage we discussed the function of ASR in section *B*.

B. Automatic Speech Recognition

ASR can be defined as the independent, computer driven transcription of spoken language into readable text in real time. This method begins once a speaker decides what to mention and really speaks a sentence. Then Software produces a speech wave kind that embodies the words of the sentence yet because the extraneous sounds and pauses within the spoken input. Next, the software makes an attempt to decrypt the speech into the simplest estimate of the sentence. Firstly it converts the speech signal into a sequence of vectors that are measured throughout the period of the speech signal. Then, employing a syntactical decoder it generates a legitimate sequence of representations.

C. Diverse Keyword Extraction

The benefit of *diverse keyword extraction* is that the coverage of the main topics of the conversation fragment is maximized. So the proposed algorithm will select a smaller number of keywords from each topic.We propose to build a topical representation of a conversation fragment, and then select content words as keywords by using topical similarity, while also rewarding the coverage of a diverse range of topics, inspired by recent summarization methods[16][17][19][20].

We aim at extracting a relevant and diverse set of keywords, cluster them into topic-specific queries ranked by importance, and present users a sample of results from these queries. The topic-based clustering decreases the chances of including ASR errors into the queries, and the diversity of keywords increases the chances that at least one of the recommended documents answers a need for information, or can lead to a useful document when following its hyperlinks.

The proposed method for diverse keyword extraction proceeds in three steps, represented schematically in Fig. 2 First, a topic model is used to represent the distribution of the abstract topic z for each word w noted as $p(z|w)$ depicted in Fig. 2. The abstract topics are not pre-defined manually but are represented by latent variables using a generative topic modeling technique. These topics occur in a collection of documents–preferably, one that is representative of the domain of the conversations. Second, these topic models are used to determine weights for the abstract topics in each conversation fragment represented by β_z . Finally, the keyword list W={w_1,...,w_k} which covers a maximum number of the most important topics are selected by rewarding diversity, using an original algorithm introduced.

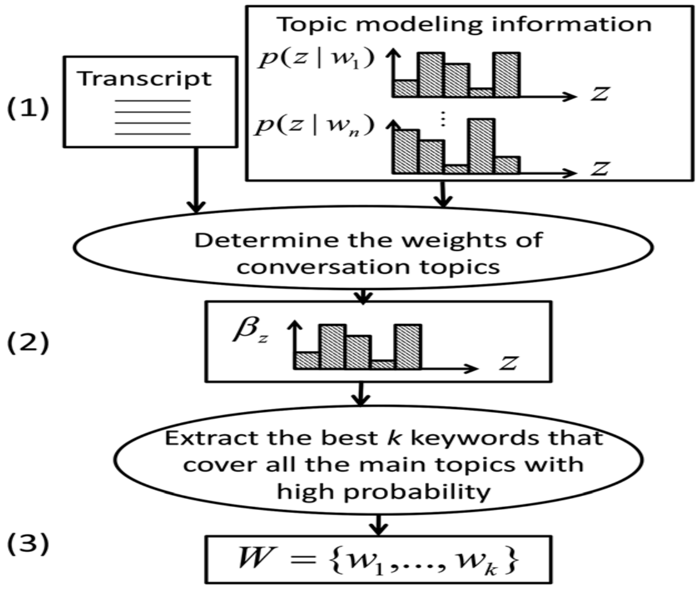

Fig. 2 The three steps of the proposed keyword extraction method

D. Clustering and Ranking

To maintain the diversity of topics embodied in the keyword set, and to reduce the ASR errors, this set must be split into several topically-disjoint subsets. Each subset corresponds then to an implicit query that will be sent to a document retrieval system. Clusters of keywords are built by ranking keywords for each main topic of the fragment. The keywords are ordered for each topic by decreasing values of $\beta_z \cdot p(z|w)$. Note that a given keyword can appear in more than one cluster. Following this ordering criterion, keywords with high value of $p(z|w)$ (i.e. more representative of the topic) will be ranked higher in the cluster of topic and these keywords will be selected from the topics with high value of β_z. Afterward, clusters themselves are ranked based on their β_z values.

E. Document Recommendations

The highest ranked keyword is sent to the just-in-time retrieval system. The system search over English Wikipedia to get the corresponding document. The most relevant retrieved document is recommended to the participants.

IV. Results and Discussion

A. Results

The input contain the oral communication of participants, that are constructed in the background from the pronounced words, obtained through real-time automatic speech recognition(ASR).The output is the most relevant document that are recommended to the participants . The document is like a Wikipedia Document with all its unwanted tags is removed and also can download the document. The valid participants can access the previous documents with a searching keyword, related to the past meeting transcript. The users can also download the document. The output is given to users with corresponding rank of the searching keyword.

B. Performance Analysis

This section, analyse the overall performance of the proposed system by computing the performance score for each module in the system. Then plot the performance graph with time. Fig 3 shows the performance analysis graph.

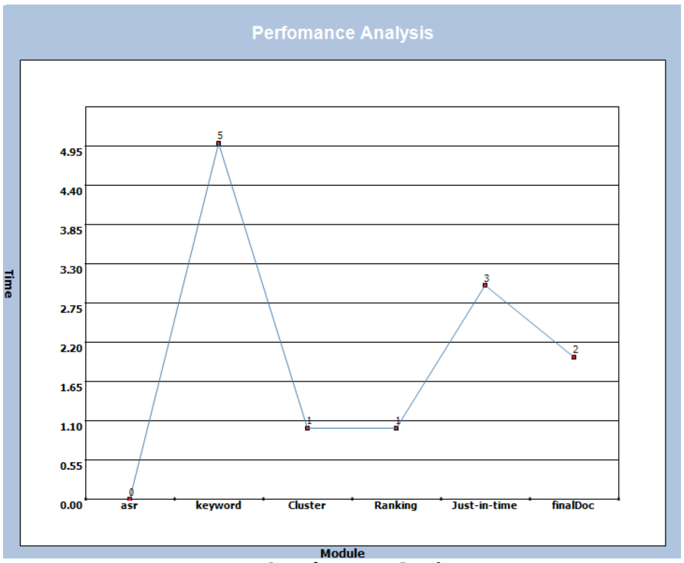

Fig 3 Performance Graph

V. Conclusion

We focused on modelling the users' information needs by deriving implicit queries from short conversation fragments. These queries are based on sets of keywords extracted from the conversation. We have proposed a novel diverse keyword extraction technique which covers the maximal number of important topics in a fragment. Then, proposed a clustering technique to divide the set of keywords into smaller topically-independent subsets constituting implicit queries.

Our current goals are to process also explicit queries, and to rank document results with the objective of maximizing the coverage of all the information needs, while minimizing redundancy in a short list of documents. Integrating these techniques in a working prototype should help users to find valuable documents immediately and effortlessly, without interrupting the conversation flow, thus ensuring the usability of our system.

ACKNOWLEDGEMENT

We thank Bodhi Info Solutions and the Idiap Social Computing group for access to the ELEA Corpus. We also acknowledge the anonymous reviewers for their precise comments and insightful remarks that improved the quality and clarity of our submission.

References

1. M. Habibi and A. Popescu-Belis, "Keyword Extraction and Clustering for Document Recommendation in Conversations" IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,VOL. 23,NO. 4, APRIL 2015.
2. M. Habibi and A. Popescu-Belis, "Enforcing topic diversity in a document recommender for conversations," in Proc. 25th Int. Conf.Comput.Linguist. (Coling), 2014, p588–599.
3. M. Habibi and A. Popescu-Belis, "Diverse keyword extraction from conversations," in Proc. 51st Annu. Meeting Assoc. Comput. Linguist.,2013, pp. 651–657.
4. D. Harwath and T. J. Hazen, "Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech,"in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012,pp. 5073–5076.
5. F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in Proc. Annu. Conf. North Amer. Chap.ACL(HLT-NAACL),2009,pp.620–628.
6. P. N. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiát, D.Korchagin, M. Lincoln, V. Wan, and L. Zhang, "Real-time ASR from meetings,"in Proc. Interspeech, 2009, pp. 2119–2122.
7. C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and Wang,"Automatic keyword extraction from documents using conditional random fields,"J. Comput. Inf. Syst., vol. 4, no. 3, pp. 1169–1180, 2008.
8. Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," Int. J. Artif.Intell. Tools, vol. 13, no. 1, pp. 157–169, 2004.
9. A. Popescu-Belis, M. Yazdani, A. Nanchen, and P. N.Garner, "A speech-based just-in-time retrieval system using semantic search," in Proc. Annu. Conf. North Amer. Chap. ACL (HLT-NAACL), 2011, pp. 80–85.
10. A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S.Castronovo, T.Wilson, A. Jaimes, and J. Carletta, "The AMIDA automatic content linking device: Just-in-time document retrieval in meetings," in Proc.5th Workshop Mach. Learn. Multimodal Interact. (MLMI), 2008, pp.272–283.
11. J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access," in Proc. 5th Int. Conf. Intell. User Interfaces (IUI'00), 2000, pp. 44–51.
12. P. E. Hart and J. Graham, "Query-free information retrieval," Int.J. Intell. Syst. Technol. Applicat., vol. 12, no. 5, pp. 32–37, 1997
13. B. J. Rhodes and P. Maes, "Just-in-time information retrieval agents,"IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.
14. M. Habibi and A. Popescu-Belis, "Using crowdsourcing to compare document recommendation strategies for conversations," Workshop Recommendat. Utility Eval.: Beyond RMSE (RUE'11), pp. 15–20,2012.
15. S.E. Cutrell, R. Sarin, and E. Horvitz, "Implicit queries (IQ)for contextualized search," in Proc. 27th Annu. Int. ACM SIGIR Conf.Res. Develop. Inf. Retrieval, 2004, pp. 594–594.
16. A.Nenkova and K. McKeown, "A survey of textsummarization techniques,"in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. New York, NY, USA: Springer, 2012, ch. 3, pp. 43–76.
17. K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A keyphrase based approach to interactive meeting summarization," in Proc. IEEE Spoken Lang. Technol. Workshop (SLT'08), 2008, pp. 153–156..
18. A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in Proc. Conf. Empir. Meth. Nat. Lang. Process.(EMNLP'03), 2003, pp. 216–223.
19. H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in Proc. 49th Annu. Meeting Assoc. Comput. Linguist.(ACL), Portland, OR, USA, 2011, pp. 510–520.
20. J. Li, L. Li, and T. Li, "Multi-document summarization via submodularity,"Appl. Intell., vol. 37, no. 3, pp. 420–430, 2012.

AUTHOR(S) PROFILE

Divya R K, Pursuing Master of Engineering in CSE at Calicut University received the Bachelor of Engineering in CSE under CUSAT and Diploma in Computer Engineering from Kerala Govt. Polytechnic College in 2008 and 2005, respectively.