IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Review on Extraction of Keywords and Recommendation of Documents in Conversation

Pallavi Gopal Patil [1], Prashant Yawalkar[2]

PG Student, Computer Engineering, MET BKC Adgaon, Nashik, SavitribaiPhule Pune University, Maharashtra, India[1]

Professor, Computer Engineering, MET BKC Adgaon, Nashik, SavitribaiPhule Pune University, Maharashtra, India[2]

Abstract: The problem of keyword extraction from conversations is with the intention to find the potentially relevant documents using the retrieval of keywords for each short conversational fragment then which are going to be recommended to participants. Short conversational fragment contains variety of words which are related to several topics; it is difficult to find out the information needed by the participants. To solve this, an algorithm to extract keywords from output of an ASR system is introduced which uses topic modeling techniques and sub modular reward function which favors diversity in keyword sets to reduce ASR noise. This method derives multiple topically separated queries from the keyword set which helps in maximizing the chances of, at least one relevant recommendation while searching these queries over English Wikipedia.

Keywords: Document recommendation, keyword extraction, topic modeling.

I. INTRODUCTION

We all are surrounded by wealth of information which is available in the form of databases, documents, or multimedia resources. But even this availability, access to this is conditioned by the availability of search engines. Users do not start searching for the information because their current activity does not allow them to do the search or they are not aware that the related information is available.

To solve this problem just in time retrieval system is adopted, which spontaneously recommend documents that are related to the current users activities. When these activities are mainly conversational, for instance when users participate in a meeting, their information needs can be modeled as implicit queries that are constructed in the background from the pronounced words, obtained through real-time automatic speech recognition (ASR). These implicit queries are used to retrieve and recommend documents from the web or local repository [15].

The goal of keyword extraction from conversations is to provide a set of words that are representative of the semantic content of the conversation. Therefore the aim is to find set of keywords, clustering of keywords and present result of this query to users in the form of documents. Mainly topic-based clustering is used to reduce the chances of including ASR errors into the queries. The focus of this is on formulating implicit queries to a just-in-time-retrieval system for use in meeting rooms. It is important that the keyword set preserves the diversity of topics from the conversation.

While the first keyword extraction methods ignored topicality as they were based on word frequencies, more recent methods have considered topic modeling factors for keyword extraction, but without specifically setting a topic diversity constraint, which is important for naturally occurring conversations [13].

Consider scenario of meeting where documents related to meeting discussion are already informed to participants of meeting. Due to some of the reasons participants does not have sufficient time to search that contents on the internet or on any other source of information. During meeting to find information related to some point is very difficult without interrupting the discussion flow. This problem occurs most of the time in meeting. To fulfill the information needs of participants some systems must be developed which will take conversation as query and give related documents to that without the direct interaction of participants to the system.

Relevance and diversity of documents can be modeled at three levels:

- While extracting queries
- Building one or several implicit queries
- Re-ranking the results of queries[14]

II. RELATED WORK

Just-in-time retrieval systems have the potential to bring advance in the process of query-based information retrieval. These systems continuously observe users' activities to detect information needs, and pro-actively retrieve relevant information. To achieve this, the systems generally extract implicit queries from the words that are written or spoken by users during their activities [15].In B some of previous keyword extraction techniques from a transcript or text are discussed.

A. *Just-in-Time Retrieval Systems*

I. Remembrance Agent:

The remembrance agent [3] performs continuous searches for information that may be relevant to user's current context. It runs continuously without user intervention. The remembrance agent uses wasted CPU cycles constructively by performing continuous searches for

DOI 10.17148/IJARCCE.2015.41204

information that might be of use in its user's current situation. RAhelps the user to organize their own data files into new categories, and helps to create continuous brainstorming session where new ideas and possible connections were suggested. The front-end of RA runsunder Emacs-19, a UNIX based text editor in elisp. The front end displays one-line suggestions along with rating indicating how relevant the documents were. The back-end is a program which, given a query-text, produces suggestions of similar documents from a pool of documents which are pre-indexed. However, since the RA runs continuously, suggestions could quickly distract from the user's primary task if they attracted too much attention. For these reasons the RA's suggestions are kept unobtrusive [2], [5].

II. Information Management Assistants system:

The architecture for Information Management Assistants system which observes the user interactions with everyday applications then anticipate their information needs through internet information sources. Information Management Assistants system fulfill information needs by using the document text the user is manipulating and knowledge of query formulation in traditional information retrieval systems. IMAs infrastructure provide information to users without requiring explicit requests. Watson system platform is used for working of this IMA's. Watson system platform is used for working of this IMA's. Watson system gathers contextual information in the form of text of the documents, which the user is manipulating in order to retrieve documents from distributed information repositories [4]. IMAs provide a framework to address the problems associated with processing queries out of context. In this the queries were constructed from the ASR using several variants of TFIDF weighting, and considering also the previous queries made by the system. The limitation of this system is that it does not incorporate semantic knowledge of particular task [6].

III. Automatic Content Linking Device (ACLD):

It is just-in-time retrieval system that constantly retrieves items from repository & displays them to participants. The overall architecture of ACLD contains four sections such as Document Bank Creator (DBC) gather documents that are of potential interest for an upcoming meeting, Document Indexer (DI) which creates an index over the document bank of the current meeting, Query Aggregator(QA) to perform document searching at regular intervals, that use words and terms which are recognized automatically from the meeting discussion and also produce a list of document names, ordered by relevance, based on the search results and on a persistence model, User Interfaces(UI) for displaying results from the QA and offers quick access to TXT/HTML and source versions of documents. AMIDA has drawbacks such as Graphical layout of user interface, Document Repository, additional functionalities such as Detecting similarities between previous discussions and current discussion would help in alerting users that they already had this discussion before [8],[11].

B. Keyword Extraction Methods

I. Keyword Extraction from a Single Document using Word Co-occurrence:

This keyword extraction algorithm is applied to a single document without using a whole database. First all frequent terms were extracted and then a set is prepared which contains co-occurrences between each term and the frequent terms. Co-occurrence distribution shows importance of a term in the document. If the term `a' is likely to be a keyword then probability distribution of co-occurrence between term `a' and the frequent terms is biased to a particular subset of frequent terms. The degree of bias of a distribution is measured by the χ^2–measure. The main advantage of this method is its simplicity without requiring use of a corpus and its high performance comparable to tfidf [1]. The disadvantage of this method is that it can only be used for single document [7].

II. Document Summarization Techniques:

D. Harwath and T. J. Hazen compared different document summarization techniques such as Direct Modeling with LexRank Extraction of 3 best frames which uses the weighted LexRank algorithm. LexRank constructs a graphical representation of a document. 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. A Maximum Marginal Relevance (MMR) based re-ranking scheme is used, which helps to preventoverlapping frames from being extracted together. This system uses a window size of 15 words, PLSA Modeling with LexRank Extraction of 3 best frames uses a different, unsupervised approach by incorporating a latent topic model into the LexRank algorithm, PLSA Modeling with LexRank Extraction of best frames assume that rather than extracting the top 3 scoring frames to compose the summary, the single frame with the highest LexRank score is extracted, Latent topic Modeling with signature word extraction extracts and displays only the top 10 unique words that are most topically relevant to the document. Extrinsic evaluation paradigm is used. Extrinsic metrics measures the utility of summarization method. These techniques are 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 [12].

III. Keyphrase Extraction:

Existing graph-based ranking methods for keyphrase extraction compute a single importance score for each word via a single random walk. Inspired by the fact that both documents and words can be represented by a mixture of semantic topics, to decompose traditional random walk into o multiple random walks specific to various topics is used. For this reason a Topical PageRank (TPR) on word graph to measure word importance with respect to different topics is build. After that, given the topic distribution of the document, the ranking scores of words are calculated and extract the top ranked ones as

DOI 10.17148/IJARCCE.2015.41204

keyphrases. Experimental results show that TPR outperforms state-of-the-art keyphrase extraction methods on two datasets under various evaluation metrics. This method does not consider topic information in other graph-based ranking algorithms such as HITS [10].

IV. Conditional Random Fields (CRF) model:

Manual assignment of high quality keywords is expensive, time-consuming, and error prone. Therefore, most algorithms and systems help people perform automatic keywords extraction. Conditional Random Fields (CRF) model is a state-of-the-art sequence labeling method, which can use the features of documents. So keywords extraction based on CRF is presented. Before keyword extraction tagging of sentences is done to find the keywords using CRF model. Processof the CRF-based Keyword Extratcion includes steps like Preprocessing and features extraction, CRF model training, CRF labeling and keyword extraction and Results evaluation. Experimental results show that the CRF model outperforms other machine learning methods such as support vector machine; multiple linear regression models etc. in the task of keywords extraction but it does not take account of the ambiguity of the extracted keywords [9].

These findings motivated to design an innovative keyword extraction method for modeling user's information needs from conversations. As mentioned in the introduction, since even short conversation fragments include words potentially related to several topics, and the ASR transcript adds additional ambiguities, a poor keyword selection method leads to non-informative queries, which often fail to capture users information needs, thus leading to low recommendation relevance and user satisfaction. So, M. Habibi and A. Popescu-Belis, introduce a novel keyword extraction technique from conversation, which maximizes the coverage of potential information needs of users, these keywords are clustered to build several topically-separated queries, results of these queries are merged into ranked set and finally these results are shown to user as recommendations. The main aim behind this system is to present recommendations related to user's current activity. The results are provided to users without initiation of direct search [15].

III. OVERVIEW OF SYSTEM

A two-stage approach is used 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. The second stage is the clustering of the keyword set in the form of several topically-disjoint queries. Figure 1 shows the system architecture consisting of main blocks such as Diverse Keyword Extraction, Keyword Clustering, and Recommendation of Documents.

There are two algorithms for extraction of keywords from text and to cluster those keywords into topic specific queries. For extraction of keywords from text diverse keywordextraction technique is used [13]. Diverse Keyword Extraction proceeds in two steps i.e., to build a topical representation of a conversation fragment, and then to select keywords using topical similarity while also rewarding the diversity of topic coverage.

To maintain the diversity of topics embodied in the keyword set, and to reduce the noisy effect of each information need on the others, 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. To improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment, with the keywords of each cluster.

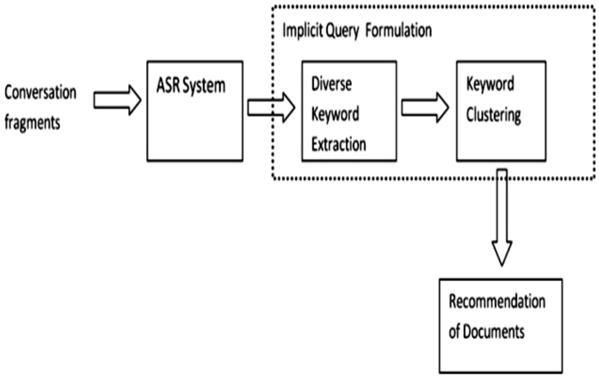
Implicit Query Formulation
Conversation fragments
ASR System
Diverse Keyword Extraction
Keyword Clustering
Recommendation of Documents

Figure 1: System Architecture

IV. CONCLUSION

Just-in-time retrieval system is considered for conversational environments which is used to recommend the relevant documents to participants as per there information need. The user's needs are modeled by formulating the implicit queries for conversation fragment. These queries are formulated by using the extracted keywords.

The novel diverse keyword extraction technique is used for finding the keywords. This technique covers the maximal number of important topics in a fragment. This keyword extraction technique is compared with techniques which uses TFIDF values and word frequencies for keyword extraction. These approaches do not consider word meaning because of this they may ignore low frequency words which together indicate highly-salient topic. Various experiments show that diverse keyword extraction technique on average the most representative keyword set. Then a clustering technique is used to divide the set of keywords into smaller topically-independent subsets constituting implicit queries.

Finally the documents are recommended to users as per their needs. As clustering technique divide the set of keywords into smaller topically-independent subsets so that it covers more topics mentioned in conversation and may help in improving the recommendations.

ACKNOWLEDGEMENT

The author is thankful to MET's Institute of Engineering Bhujbal Knowledge City Nasik, HOD of computer department, guide and parents for their blessing, support and motivation behind this work**.**

DOI 10.17148/IJARCCE.2015.41204

REFERENCES

[1] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” In Information Processing & Management. Journal, vol. 24, no. 5, pp. 513–523, 1988.

[2] B. Rhodes and T. Starner, “Remembrance Agent: A continuously running automatedinformation retrieval system”, in Proc. 1st Int. Conf. Pract. Applicat. Intell.Agents Multi Agent Technol., London, U.K., 1996, pp. 487-495.

[3] B. J. Rhodes, “The wearable Remembrance Agent: A system for augmented memory,” InPersonal Technol., vol. 1, no. 4, pp. 18–224, 1997.

[4] Budzik, J., and Hammond, K. “Watson: Anticipating and Contextualizing Information Needs”, In Proceedings of the ASIS 1999 Annual Conference. Information Today, Inc., Medford NJ, 1999.

[5] B. J. Rhodes and P. Maes, “Just-in-time information retrieval agents,” In IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.

[6] A. J. Budzik and K. J. Hammond, “User interactions with everyday applicationsas context for just-in-time information access”, in Proc. 5th Int. Conf. Intell.User Interfaces (IUI00), 2000, pp. 44-51.

[7] Y. Matsuo and M. Ishizuka,”Keyword extraction from a single document using word co-occurrence statistical information”, in Int. J. Artif. Intell. Tools, vol. 13, no. 1, pp. 157-169, 2004.

[8] 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-timedocument retrieval in meetings”, in Proc. 5th Workshop Mach. Learn.Multimodal Interact. (MLMI), 2008, pp. 272-283.

[9] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, “Automatic keyword extraction from documents using conditional random fields,” *J. Comput. Inf. Syst.*, vol. 4, no. 3, pp. 1169–1180, 2008.

[10] Z. Liu, W. Huang, Y. Zheng, and M. Sun, “Automatic keyphrase extraction via topic decomposition,” in Proc. Conf. Empir. Meth. Nat. Lang. Process. (EMNLP’10), 2010, pp. 366–376.

[11] 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. NorthAmer. Chap. ACL (HLT-NAACL), 2011, pp. 80-85.

[12] D. Harwath and T. J. Hazen, “Topic identification based extrinsic evaluation ofsummarization techniques applied to conversational speech”, in Proc. Int. Conf.Acoust., Speech, Signal Process. (ICASSP), 2012, pp. 5073-5076.

[13] M. Habibi and A. Popescu-Belis, “Diverse keyword extraction from conversations”, inProc. 51st Annu. Meeting Assoc. Comput. Linguist. 2013, pp. 651-657.

[14] M. Habibi and A. Popescu-Belis, “Enforcing topic diversity in a document recommenderfor conversations”, in Proc. 25th Int. Conf. Comput. Linguist. (Coling), 2014, pp. 588-599.

[15] M. Habibi and A. Popescu-Belis, “Keyword Extraction and Clustering for Document Recommendation in Conversations”, in IEEE transaction on Audio, Speech, and language processing, Vol. 23, No. 4, 2015, pp. 746-759.

DOI 10.17148/IJARCCE.2015.41204