ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 2, pp: (15-21), Month: April - June 2016, Available at: www.researchpublish.com

Extraction of Keywords from Conversation for Recommending Document

[1]Hitesh Purandare, [2]Abhishek Singh, [3]Niket Vyas, [4]Chhagan Chavan

International Institute of Information Technology, Pune, India

Abstract: **This paper addresses the issue of keyword extraction from conversations, with the objective of utilizing these watchwords to recover, for every short discussion piece, a little number of conceivably pertinent reports, which can be prescribed to members. In any case, even a short piece contains a mixed bag of words, which are conceivably identified with a few themes; also, utilizing a programmed discourse acknowledgment (ASR) framework presents slips among them. Along these lines, it is hard to surmise correctly the data needs of the discussion members. We first propose a calculation to remove decisive words from the yield of an ASR framework (or a manual transcript for testing), which makes utilization of theme demonstrating methods and of a sub modular prize capacity which supports differing qualities in the magic word set, to coordinate the potential differing qualities of subjects and decrease ASR commotion. At that point, we propose a technique to infer various topically isolated inquiries from this decisive word set, keeping in mind the end goal to amplify the possibilities of making at any rate one pertinent proposal when utilizing these questions to seek over the English Wikipedia. The proposed systems are assessed as far as significance as for discussion pieces from the Fisher, AMI, and ELEA conversational corpora, appraised by a few human judges. The scores demonstrate that our proposition moves forward over past systems that consider just word recurrence or theme closeness, and speaks to a promising answer for a report recommender framework to be utilized as a part of discussions.**

Keywords: **Document recommendation, information retrieval, keyword extraction, meeting analysis, topic modelling**.

I. INTRODUCTION

Humans are encompassed by an uncommon abundance of data, accessible as records, databases, or mixed media assets. Access to this data is adapted by the accessibility of suitable web indexes, however notwithstanding when these are accessible, clients frequently don't start a pursuit, in light of the fact that their current action does not permit them to do as such, or in light of the fact that they are not mindful that applicable data is accessible. We receive in this paper the point of view of in the nick of time recovery, which replies this inadequacy by suddenly suggesting archives that are identified with clients' present exercises. At the point when these exercises are primarily conversational, for occurrence when clients take part in a meeting, their data needs can be demonstrated as understood inquiries that are built out of sight from the professed words, acquired through continuous programmed discourse acknowledgment (ASR). These certain questions are utilized to recover and suggest reports from the Web or a neighbourhood storehouse, which clients can decide to, investigate in more detail if they discover them intriguing. The centre of this paper is on figuring verifiable questions to a without a moment to spare recovery framework for utilization in meeting rooms. Conversely to unequivocal talked inquiries that can be made in business Web crawlers, our in the nick of time recovery framework must develop certain questions from conversational information, which contains a much bigger number of words than a question. For example, in the illustration examined in Section V-B underneath, in which four individuals set up together a rundown of things to help them get by in the mountains, a short piece of 120 seconds contains around 250 words, relating to a mixed bag of areas, for example, 'chocolate', 'gun', or 'lighter'. What might then be the most supportive 3–5 Wikipedia pages to prescribe, and how might a framework focus them? Given the potential variety of themes, strengthened by potential ASR

slips or discourse disfluencies, (for example, "rush" in this illustration), our objective is to keep up different speculations about clients' data needs, and to present a little example of proposals in view of the no doubt ones. In this manner, we point at separating a pertinent and various arrangements of catchphrases, group them into theme particular questions positioned by significance, and present clients an example of results from these questions. The point based bunching abatements the possibilities of including ASR blunders into the questions, and the assorted qualities of essential words expands the possibilities that no less than one of the suggested records answers a need for data, or can prompt a helpful archive while taking after its hyperlinks. Case in point, while a strategy in view of word recurrence would recover the accompanying Wikipedia pages: 'Light', 'Lighting', and 'Light My Fire' for the aforementioned piece, clients would lean toward a set, for example, 'Lighter', "Fleece" and 'Chocolate'. Pertinence and assorted qualities can be authorized at three stages: at the point when removing the magic words; when building one or a few certain inquiries; or when re-positioning their outcomes.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

We addressed the problem of building concise, diverse and relevant lists of documents, 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. We propose in this paper an algorithm for diverse merging of these lists, using a sub modular reward function that rewards the topical similarity of documents to the conversation words as well as their diversity. We evaluate the proposed method through crowdsourcing. The results show the superiority of the diverse merging technique over several others which not enforce the diversity of topics.

In other paper info is that the statistical approach to mechanized encoding and searching of literary information communication of ideas is carried out on the basis of statistical probability in that a writer chooses that level of subject specificity and that combination of words which he feels will convey the most meaning. Since this process varies among individuals and since similar ideas are therefore relayed at different levels of specificity and by means of different words, the problem of literature searching by machines still presents major difficulties. A statistical approach to this problem will be outlined and the various steps of a system based on this approach will be described. Steps include the statistical analysis of a collection of documents in a field of interest, the establishment of a set of “notions” and the vocabulary by which they are expressed, the compilation of a thesaurus-type dictionary and index, the automatic encoding of documents by machine with the aid of such a dictionary, the encoding of topological notations (such as branched structures), the recording of the coded information, the establishment of a searching pattern for finding pertinent information, and the programming of appropriate machines to carry out a search.

We argue that the quality of a summary can be evaluated based on how many concepts in the original document(s) that can be preserved after summarization. Here, a concept refers to an abstract or concrete entity or its action often expressed by diverse terms in text. Summary generation can thus be considered as an optimization problem of selecting a set of sentences with minimal answer loss. In this paper, we propose a document concept lattice that indexes the hierarchy of local topics tied to a set of frequent concepts and the corresponding sentences containing these topics. The local topics will specify the promising sub-spaces related to the selected concepts and sentences. Based on this lattice, the summary is an optimized selection of a set of distinct and salient local topics that lead to maximal coverage of concepts with the given number of sentences. Our summarizer based on the concept lattice has demonstrated competitive performance in Document Understanding Conference 2005 and 2006 evaluations as well as follow-on tests.

***Linking educational materials to encyclopaedic knowledge**:*

This paper describes a system that automatically links study materials to encyclopaedic knowledge, and shows how the availability of such knowledge within easy reach of the learner can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge.

Remembrance Agent:

The Remembrance Agent (RA) is a program which augments human memory by displaying a list of documents which might be relevant to the user’s current context. Unlike most information retrieval systems, the RA runs continuously without user intervention. Its unobtrusive interface allows a user to pursue or ignore the RA’s suggestions as desired.

III. OUR STUDIES AND FINDINGS

ASR: Speech recognition (SR) is the inter-disciplinary subfield of computational linguistics which incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields to develop methodologies and technologies that enables the recognition and translation of spoken language into text by computers and computerized devices such as those categorized as Smart Technologies and robotics. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text". Speech recognition applications include voice user interfaces such as voice dialling (e.g. "Call home"), call routing (e.g. "I would like to make a collect call"), demotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing. The ultimate goal of ASR is research is to allow a computer to recognize in real time, with 100% accuracy, all words that are intelligibly spoken by a person, independent by vocabulary size, noise, speaker characteristics or accent. Today if the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible and accuracy can be greater than 90%. This explains why some users, especially those whose speech is heavily accented, might achieve recognition rates much lower than expected. You speak to the software via an audio feed The device you're speaking to creates a wave file of your words:

1) The wave file is cleaned by removing background noise and normalizing volume

2) The resulting filtered wave form is then broken down into what are called phonemes. (Phonemes are the basic building block sounds of language and words.

3) English has 44 of them, consisting of sound blocks such as "wh", "th", "ka" and "t".

4) Each phoneme is like a chain link and by analyzing them in sequence, starting from the first phoneme, the ASR software uses statistical probability analysis to deduce whole words and then from there, complete sentences.

5) Your ASR, now having "understood" your words, can respond to you in a meaningful way.

Keyword Extraction:

Various strategies have been proposed to consequently remove pivotal words from a content, and are relevant additionally to interpreted discussions. The most punctual procedures have utilized word frequencies and TFIDF qualities to rank words for extraction. On the other hand, words have been positioned by checking pairwise word co-event frequencies. These methodologies don't consider word significance, so they may overlook low-recurrence words which together demonstrate an exceedingly notable subject. For example, the words 'auto', 'wheel', 'seat', and "traveller" happening together demonstrate that autos are a notable theme regardless of the fact that every word is not itself incessant. To enhance over recurrence based strategies, a few approaches to utilize lexical semantic data have been proposed. Semantic relations between words can be acquired from a physically developed thesaurus, for example, Word Net, or from Wikipedia, or from a naturally assembled thesaurus utilizing idle subject displaying strategies, for example, LSA, PLSA, or LDA. For example, pivotal word extraction has utilized the recurrence of all words having a place with the same WorldNet idea set, while the Wikifier framework depended on Wikipedia connections to register another substitute to word recurrence. Hazen also applied topic modelling techniques to audio files. In another study, he used PLSA to build a thesaurus, which was then used to rank the words of a conversation transcript with respect to each topic using a weighted point-wise mutual information scoring function. Moreover, Harwath and Hazen utilized PLSA to represent the topics of a transcribed conversation, and then ranked words in the transcript based on topical similarity to the topics found in the conversation. Similarly, Harwath *et al.* extracted the keywords or key phrases of an audio file by directly applying PLSA on the links among audio frames obtained using segmental dynamic time warping, and then using mutual information measure for ranking the key concepts in the form of audio file snippets. A semi-supervised latent concept classification algorithm is done using LDA topic modelling for multi-document information extraction

Diverse Keyword Extraction:

We propose to take advantage of topic modelling techniques 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. The benefit of *diverse keyword extraction* is that the coverage of the main topics of the conversation fragment is maximized. Moreover, in order to cover more topics, the proposed

algorithm will select a smaller number of keywords from each topic. This is desirable for two reasons; this will lead to more dissimilar implicit queries, thus increasing the variety of retrieved documents and if words which are in reality ASR noise can create a main topic in the fragment, then the algorithm will choose a smaller number of these noisy keywords compared to algorithms which ignore diversity.

Algorithm 1: Diverse keyword extraction.

Input: a given text t, a set of topics Z, the number of keywords k

Output: a set of keywords S

$S \leftarrow \emptyset$

While $|S| \leq k$ **do**

$S \leftarrow S \cup \{argmax_{w \in t \setminus S}(h(w, S))$ where $h(w, S) = \sum_{z \in Z} \beta_z [p(z|w) + r_{S,z}]^{\lambda}$;

end

return S

Fig. The three steps of the proposed keyword extraction method:(1) topic modeling, (2) representation of the main topics of the transcript, and (3) diverse keyword selection.

Keyword Clustering

The diverse set of extracted keywords is considered to represent the possible information needs of the participants to a conversation, in terms of the notions and topics that are mentioned in the conversation. 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. These subsets are obtained by clustering topically-similar keywords, as follows.

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 β_z.p(z|w) Moreover, in each cluster, only the keywords with aβ_z.p(z|w) value higher than a threshold are kept for each topic .Note that a given keyword can appear in more than one cluster. Following this ordering criterion, keywords with high value of will be ranked higher in the cluster of topic and these keywords will be selected from the topics with high value of. Afterward, clusters themselves are ranked based on their values.

From Keywords to Document Recommendations:

As a first idea, one implicit query can be prepared for each conversation fragment by using as a query all keywords selected by the diverse keyword extraction technique. However, to improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment, with the keywords of each cluster from the previous section, ordered as above (because the search engine used in our system is not sensitive to word order in queries). In experiments with only one implicit query per conversation fragment, the document results corresponding to each conversation fragment were prepared by selecting the first document retrieval results of the implicit query.

JIT Retrieval:

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. We review existing just-in-time-retrieval systems and methods used by them for query formulation. A just-in-time information retrieval agent (JITIR agent) is software that proactively

retrieves and presents information based on a person's local context in an easily accessible yet nonintrusive manner. This paper describes three implemented JITIR agents: the Remembrance Agent, Margin Notes, and Jimminy. Theory and design lessons learned from these implementations are presented, drawing from behavioral psychology, information retrieval, and interface design. They are followed by evaluations and experimental results. The key lesson is that users of JITIR agents are not merely more efficient at retrieving information, but actually retrieve and use more information than they would with traditional search engines. Hence it is used and important factor in the process where extraction and clustering should be done fast as possible.

Ranking:

Clusters of keywords are built by ranking keywords for each main topic of the fragment. Afterward, clusters themselves are ranked based on their β_z values.

Selection of Configurations:

Using rank biased overlap as a similarity metric, based on the fraction of keywords overlapping at different ranks.

$$RBO(S,T) = \frac{1}{\sum_{d=1}^{D} (\frac{1}{2})^{d-1}} \sum_{d=1}^{D} (\frac{1}{2})^{d-1} \frac{|S_{1:d} \cap T_{1:d}|}{|S_{1:d} \cup T_{1:d}|}$$

Where,

RBO = rank biased overlap Sand T be two ranked lists, and Si be the keyword at rank i in S The set of the keywords up to ranked in S is {Si : I : <= d}.noted as $S_{1:d}$. RBO is calculated as above Equation

Document Recommendation:

One implicit query can be prepared for each conversation fragment by using as a query all keywords selected by the diverse keyword extraction technique. However, to improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment, with the keywords of each cluster from the previous section, ordered as above (because the search engine used in our system is not sensitive to word order in queries)

IV. CONCLUSION

We have considered a specific type of without a moment to spare recovery frameworks proposed for conversational situations, in which they prescribe to client's archives that are important to their data needs. We concentrated on displaying the client's data needs by getting verifiable questions from short discussion pieces. These questions are in light of sets of pivotal words separated from the discussion. We have proposed a novel different pivotal word extraction strategy which covers the maximal number of vital themes in a piece. At that point, the boisterous impact on questions of the blend of themes in a decisive word set, we proposed a grouping system to isolate the arrangement of catchphrases into littler topically-autonomous subsets constituting understood inquiries.

We compared the diverse keyword extraction technique with existing methods, based on word frequency or topical similarity, in terms of the representativeness of the keywords and the relevance of retrieved documents. These were judged by human ratters recruited via the Amazon Mechanical Turk crowd sourcing platform. The experiments showed that the diverse keyword extraction method provides on average the most representative keyword sets, with the highest - NDCG value, and leading–through multiple topically-separated implicit queries–to the most relevant lists of recommended documents. Therefore, enforcing both relevance and diversity brings an effective improvement to keyword extraction and document retrieval. The keyword extraction method could be improved by considering n-grams of words in addition to individual words only, but this requires some adaptation of the entire processing chain.

ACKNOWLEDGEMENT

The authors are grateful to the Hasler Foundation for its support through the REMUS project (Re-ranking Multiple Search Results for Just-in-Time Document Recommendation, 2014). The authors also thank the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] M. Habibi and A. Popescu-Belis, "Enforcing topic diversity in a document recommender for conversations," in *Proc. 25th Int. Conf. Comput. Linguist. (Coling)*, 2014, pp. 588–599.

[2] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Develop.*, vol. 1, no. 4, pp. 309–317, 1957.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage. J.*, vol. 24, no. 5, pp. 513–523, 1988.

[4] S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu, "Document concept lattice for text understanding and summarization," *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1643–1662, 2007.

[5] A. Csomai and R. Mihalcea, "Linking educational materials to encyclopedic knowledge," in *Proc. Conf. Artif. Intell. Educat.: Building Technol. Rich Learn. Contexts That Work*, 2007, pp. 557–559.

[6] 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.

[7] 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.

[8] 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.

[9] P. E. Hart and J. Graham, "Query-free information retrieval," *Int. J. Intell. Syst. Technol. Applicat.*, vol. 12, no. 5, pp. 32–37, 1997.

[10] B. Rhodes and T. Starner, "Remembrance Agent: A continuously running automated information retrieval system," in *Proc. 1st Int. Conf. Pract. Applicat. Intell. Agents Multi Agent Technol.*, London, U.K., 1996, pp. 487–495.

APPENDIX - A

Transcript of a Conversation Fragment from the ELEA Corpus

The following transcript of a conversation fragment (speakers noted A through C) was submitted to the document recommender system.

A: okay I start.

B: how how do you want to proceed?

A: I guess -

C: yes what is the most important?

A: I guess fire light.

B: fire lighter?

A: fire, yes. I would say if we had something we can fire with -- I guess that the lighter is useful in getting some sparks.

B: hopefully.

A: so we can use either newspaper or -- something like that.

C: but again - first it is more important to have enough err clothes.

A: and for me, more important to know where to go. I would say that the compass.

C: I mean -- if you don't have enough clothes so -- at one point you can --

B: you can die.

C: yes you can -- you will die. so first issue, try to keep yourself alive and then you can --

A: but -- but you already have some --

B: basics. You everything. You have enormous which is and so is no shoes here.

C: okay that we have shoes so -- okay.

B: because seventy kilometres will take you how many days? Err in the snow -- what do you think?

A: two or three.

B: it can be two or three days?

C: yes, but okay you cannot always have fire with you -- but you need always have clothes with you. I mean it is the only thing that protects you when you are walking.

B: oh yes. and arm you can make an igloo during the evening. not that cold. only about five degrees. so lighting a fire is not so important.

C: I guess fire is an extra. I mean it is important but err for me first it is important that when you keep walking you should be protected.