DOI : https://doi.org/10.56452/7-2-562

ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Document Recommendation In Conversations using Keyword Extraction And Clustering Method

Chandradeep Bhatt

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University,
Dehradun, Uttarakhand India 248002,

Abstract

In order to satisfy the information demands of discussion participants without diverting them, it is necessary to create lists of papers that are succinct, varied, and relevant. This study tackles this challenge. These lists are regularly fetched by sending several implicit requests generated from the spoken phrases. Each query is sent to a search engine over the English Wikipedia and is connected to one of the subjects mentioned in the conversation fragment before the recommendation. Utilizing a sub-modular reward function that rewards both the thematic closeness of documents to the conversation terms and their diversity, we offer an algorithm in this study for the diverse merging of these lists. Via crowdsourcing, we assess the suggested approach. The outcomes demonstrate the effectiveness of the diverse merging strategy over a number of alternatives that do not uphold subject diversity. In this study, we provide a unique method for extracting keywords from the ASR output that increases the coverage of users' prospective information demands while minimising the usage of irrelevant terms. After a collection of keywords has been extracted, they are clustered to create a number of topically separated searches that may be performed individually and provide more accuracy than a single, larger topically mixed query. Before providing the results as suggestions to users, the results are eventually combined into a ranked set.

1. INTRODUTION

The approach for combining lists of documents that were retrieved using many implicit queries that were designed for quick talks is presented in this work. The objective is to create a distinct and condensed list of papers that may be suggested to discussion participants in real time. The technique draws influence from earlier work on varied keyword extraction as well as extractive text summarization. To address the greatest number of implicit questions and themes in a brief and pertinent list of recommendations, it rewards topic variety and similarity. The technique increases coverage to its fullest potential of subjects that are automatically identified in conversation transcripts.

The approach is assessed using samples from the Fisher and AMI corpora and a crowdsourcing platform to elicit judgments about relative importance. The current state of automatic key phrase extraction is examined in this research, along with the main causes of mistakes produced by current systems and future difficulties. The findings show that the technique beats two competing baselines, although the performance on this task at the cutting edge is still much behind that on many fundamental NLP tasks. The purpose of automatic key phrase extraction is to extract a group of phrases that are linked to the primary subjects covered in a given text by automatically selecting significant as well as pertinent terms from the text's body. Document key words have made it possible to quickly and accurately search for a specific document inside a vast text collection and have shown their potential to enhance numerous information retrieval (IR) and natural language processing (NLP) operations. The challenge is still very much open, though.

This study provides a latent concept classification–based semi-supervised extractive summarization approach. It is trained using a probabilistic Bayesian model on hidden ideas found in documents and the accompanying human-written summaries. According to experimental findings, selecting sentences for inclusion in summary text based on the categorization of latent summary concepts enhances the quality of the summaries that are produced. The supervised framework and unsupervised Maximum Marginal Relevance (MMR) techniques are both examined in this research. To make the extractive summary more readable and similar to an abstractive summary, it is suggested that sentence compression be applied to it.

Moreover, it assesses how employing compressed utterances affects summary generation and suggests a completely automated summarizer that produces compressed meeting summaries. Lastly, it examines using the concept-based global optimization strategy to summarize the Twitter subjects and Twitter's non-standard tokens were converted into words using a unique letter translation technique.

Users may browse through a huge number of documents rapidly with the use of automatic keyword extraction and summarization, which has been the subject of extensive research in recent years. Emails, forums, Twitter tweets, and taped meeting dialogues are just a few examples of the non-traditional information communication technologies that have fast become important information sources. The language produced from these data sources has a conversational tone and differs significantly from printed materials. Several people often contribute to conversational text, which has a poor text coherence, a lot of repetitions, and disfluencies. Although processing informal conversational text presents significant technological obstacles, gist information may be extracted from these resources with tremendous success.

Users may browse through a huge number of documents rapidly with the use of automatic keyword extraction and summarization, which has been the subject of extensive research in recent years. Emails, forums, Twitter tweets, and taped meeting dialogues are just a few examples of the non-traditional information communication technologies that have fast become important information sources. The language produced from these data sources has a conversational tone and differs significantly from printed materials. Several people often contribute to conversational text, which has a poor text coherence, a lot of repetitions, and disfluencies. Although processing informal conversational text presents significant technological obstacles, gist information may be extracted from these resources with tremendous success.

2. LITERATURE SURVEY

This paper proposes a method for combining lists of documents obtained via various implicit queries designed for brief exchanges. It makes use of a submodular reward mechanism that awards both the variety and thematic resemblance of texts to everyday terms. We draw ideas from extractive text summarization and our own prior work on diversified keyword extraction to combine the lists of texts based on these criteria. The ACLD system is a recommender system for conversational contexts that listens to the conversation as it is happening and creates queries based on the words that an ASR system is able to recognise in real-time. The system segments the discussion at the end of the closest syllable and uses the complete conversation fragment since the last suggestion to suggest documents every two minutes. For integrating lists of documents from numerous thematically distinct implicit inquiries, we developed a variety of merging techniques. These techniques were created using keyword lists derived from the transcripts of conversational fragments [1].

This study looks at a summary's extrinsic usefulness, which is determined by how well it can help a person do a certain activity. It conducted a large-scale human study comparing four alternative summarising methods applied to conversational speech from the Fisher Corpus using Amazon's Mechanical Turk service. 152 different employees engaged in the 1374 distinct HITs that were uploaded to Mechanical Turk. The findings indicated that automatic topic identification systems can be utilised as quick, low-cost evaluation tools since the performance of the human test participants and an automated topic identification system on the same task appeared to be associated. Human subjects outperformed the machine system when context was incorporated into the summaries, however the machine system was more adapted to the usage of system 4 than the human subjects. In addition to producing the most concise summaries, system 4's use of signature key phrases also led to much higher subject ID accuracy than the other methods. The direct cosine similarity measure in system 1 was exceeded by the PLSA based frame similarity measure in system 2. This shows that adding a latent topic model to an extractive summarization system significantly improves performance compared to using a direct model [2].

An continuing discussion or monologue is monitored by the Automated Content Linking Device (ACLD), a just-in-time document retrieval system, which then enhances it with possibly relevant materials, including multimedia ones, from nearby archives or the Internet. Participants in meetings or people watching a recorded lecture or chat can see results in real time. The words retrieved from the ASR are filtered for stop words, and the system employs external search engines to look via external repositories. It makes use of pre-determined keywords to boost the prominence of certain phrases in search results, and it computes a measure of semantic similarity between text fragments using a random walk across the network of Wikipedia articles. A group at the University of Edinburgh ran a pilot test using an older design of the unobtrusive UI. Future research will focus on enhancing semantic search's relevance, modelling context to speed up the delivery of results, and deducing relevant feedback from users [3].

The efforts made to get Boston's Museum of Science's virtual tour guides Ada and Grace ready to speak with kids and other visitors are covered in this text. Data was gathered at the museum and coded for the speaker type after being transcribed (child, adult male, adult female, or no speech). Up until December 2010, the SONIC toolkit handled speech recognition; going forward, USC's O to Sense recognition engine will take over. Direct visitor engagement and combined staff and tourist interaction were the two conditions examined. In-depth interviews with visitors following their contact, monitoring of visitors while they engaged with the displays, and follow-up online questionnaires six weeks later were the three methodologies employed in the study. The number and types of questions the visitor posed to the Twins, the classification of the Twins' replies, and visits to the Science Behind display were among the observational data. Other variables were group size and composition, stay length, social interaction types, usability problems experienced when using the exhibit, and group size and types of inquiries. With the intention of gathering a paired observation and interview with the same subject, interviews were held after visitors had interacted with either display. In this study, enhancements to the Virtual Human Museum Guides were detailed, transforming the system from one that could only be exhibited by a professional to one that could communicate with visitors directly [4].

This article analyses the influence of cooperation and task type on users' query behaviour in exploratory online search. It combines two search scenarios, collaborative search and solo search, as well as two search tasks, utility-based decision-making and recall-oriented knowledge collection. The findings indicated that while queries in solitary search were more often used reconstruction, they are more diversified in collaborative search and recall-oriented activities. In addition, people that search in teams tend to utilise New and Specialization more frequently and coordinate their searches. The goal of this study was to determine whether offering query suggestions in collaborative search can increase the rate at which queries are successfully reformulated. It was discovered that offering awareness support in addition to query suggestions may assist team members in choosing suggested queries. To completely comprehend how query reformulation supports collaborative search, further research is required [5].

3. PROPOSED SYSTEM

We provide a two-stage method for creating implicit inquiries. The initial step is to extract keywords from a conversation fragment's transcript in order to propose documents using information from an ASR system. Beyond the conventionally common features, we also present rich characteristics that are speaker and prominence based, connected to decision-making sentences, and features that are taken from automatically generated summaries. We concentrate on supervised framework for keyword extraction and incorporate various knowledge sources. We provide a feedback technique to reinforce

DOI : https://doi.org/10.56452/7-2-562

the influence of summary sentences on keyword selection. We do analysis to assess feature efficacy using various feature selection methods, and we create a number of metrics to gauge the calibre of summaries that might be useful for the keyword extraction assignment.

The effectiveness of the system is further assessed using human transcripts and two types of ASR output (1-best and n-best). The results of keyword extraction using n-best ASR output over 1-best hypothesis are promisingly enhanced. We investigate many meeting-specific factors for extractive meeting summarization. To enhance the effectiveness of extractive meeting summarization, we suggest using subject labels and speaker-dependent traits (that is verbosity, gender, native language, and role in the meeting). Both the supervised framework and the unsupervised Maximum Marginal Relevance (MMR) method included these qualities. On both human transcripts and ASR output, also applying a variety of measurement statistics, such as ROUGE, Pyramid, and a DA-level F-measure score, we consistently see improvements when using our suggested approaches.

To make the extractive summary more readable and similar to an abstractive summary, we recommend doing sentence compression on it in addition to extractive summarization. Sentences might be shortened as a first step towards our eventual objective of producing an abstract for spoken materials. The noisy-channel method, the integer linear programming (ILP) based technique with filler phrase identification, the noisy-channel approach employing Markovization formulation of grammatical rules, and various other automated compression algorithms are also investigated. We use the Amazon Mechanical Turk to do large-scale speech compression annotation, and we contrast the automatically compressed utterances to human compression as well as to the abstractive summaries.

Moreover, we assess how employing compressed utterances affects summarising. We next present a completely automated summarizer that combines an extractive summary system with an utterance compression module to provide compressed meeting summaries.

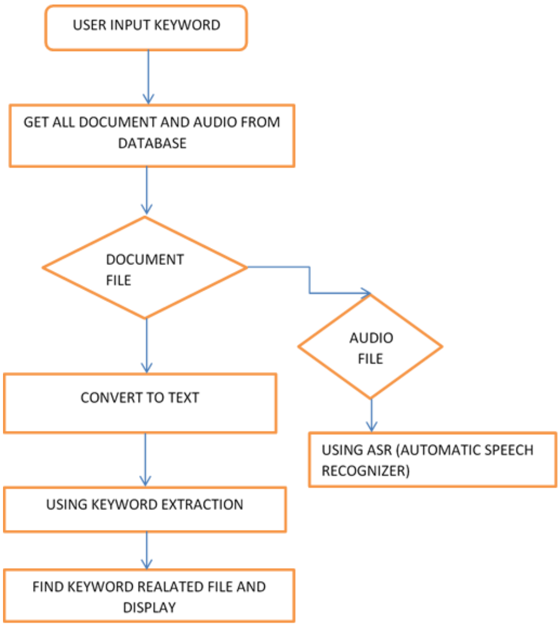

USER INPUT KEYWORD
GET ALL DOCUMENT AND AUDIO FROM DATABASE
DOCUMENT FILE
AUDIO FILE
CONVERT TO TEXT
USING ASR (AUTOMATIC SPEECH RECOGNIZER)
USING KEYWORD EXTRACTION
FIND KEYWORD REALATED FILE AND DISPLAY

Fig 1: Flow Diagram

According to our research, extractive summaries can be compressed to increase human readability and the ROUGE scores when compared to the original, uncompressed extractive summaries.

There are numerous approaches to exploit lexical semantic information that outperform frequency-based techniques. A manually compiled thesaurus, such as Word Net, Wikipedia, or an automatically compiled thesaurus created using latent topic modelling approaches, such as LSA, PLSA, or LDA, can all provide information on the semantic relationships between words. The ELEA Corpus (Emergent Leader Analysis) is made up of about ten hours' worth of meetings that were recorded and then transcribed in both English and French. Participants must rate a list of 12 objects according to their utility for living on the mountain until they are rescued in a role-playing game that takes place at each meeting. For our studies, we used 5 English conversations that lasted around 15 minutes each. We then broke the transcripts into 35 parts that lasted approximately 2 minutes each. The transcripts in the ELEA Corpus, like the AMI Corpus, were insufficient for topic modelling. Therefore, topic models were learned using the same subset of the English Wikipedia that was used for the AMI Corpus.

Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) are analytical methods for investigating relationships between data points. Assuming that words with the same or similar meanings are frequently employed in the same or similar settings, LSI is based on the mathematical approach of Singular Value Decomposition. Using their contextual occurrences, the strategy is utilised to create linkages between previously undiscovered or unknown words. The approaches described in this part illustrate how natural language processing skills may be applied, although they typically do not provide a complete answer to the analytical problem at hand. The final part, application areas, makes use of these strategies.

The following are some of the proposed approach's benefits:

- The frequency of all terms in the same WordNet concept set was utilised for keyword extraction.
- Utilizing a weighted point-wise mutual information scoring mechanism, a thesaurus was created using PLSA and used to rate the words in a conversation transcript with regard to each Subject.
- Models for key word extraction have been learned using supervised machine learning techniques.
- It served as a gauge of the keyword list's topic variety.

1. Conversation to text document

Get dialogue from meetings and chats. The chat should then be converted to a text document. The benifit of diversified keyword extraction is that it maximises the coverage of the conversation fragment's primary subjects. Also, the suggested method will choose fewer keywords from each subject in order to cover more topics. For two reasons, this is beneficial. Secondly, the keyword clustering will result in more diverse implicit searches, broadening the range of documents returned. Second, compared to algorithms that ignore variety, the algorithm will choose fewer of the words that are ASR noise if they can establish a core theme in the fragment.

2. Get text conversation

Building a Java frame, then retrieving all text conversations from a text file that has been saved. We compared the three document sets produced by the D (.75), TS, and WF keyword extraction methods using single queries built from the keyword sets they provided. Second, we created several searches using the same techniques and made comparable comparisons across the document sets that resulted. Lastly, we contrasted the top outcomes of single-query and multi-query searches. Due to the time-consuming nature of evaluations involving human beings, we tried to carry out as few comparisons as possible while still enabling the techniques to be ordered.

3. Reduction noise

This technique cuts down on generic words in text documents to lessen the impact of noise in meeting settings. It conducted binary comparisons between the keyword lists derived using four crowd-sourced key-word extraction techniques over the eight pieces from the real-time ASR system's ASR transcripts of the AMI Corpus. The mean relevance values for all comparisons required to rank these approaches are displayed, and D (.75) continues to beat TS and WF D. (.5). Even with ASR noise present, the ranking does not change.

4. Retrieving relevant document

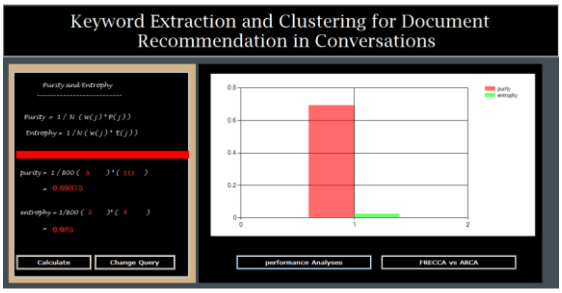
Discovering the document's keyword, followed by similar documents. Several inquiries Compared to single queries (TS and D(.75)), CTS and CD(.75)) retrieve a significant number of pertinent documents. This is likely because single queries cannot distinguish between the variety of conversational topics and produce irrelevant results like "Shorten," "Whisk," and "Fly-whisk" (found by TS) and "25metre rapid

DOI : https://doi.org/10.56452/7-2-562

fire pistol," "Fire safe cigarettes," and "25metre center-fire pistol" (found by D(.75)). Also, compared to CTS, CD (.75) discovers papers that cover more of the themes stated in the conversational segment.

4. RESULTS

This essay looks at the issue of creating succinct, varied, and pertinent lists of publications to suggest to conversation participants. It also provides a unique keyword extraction approach, looks at several automatic compression techniques, and suggests an algorithm for the varied merging of these lists. It also uses Amazon Mechanical Turk to carry out extensive utterance compression annotation. Last but not least, it suggests a completely automated summarizer that creates compressed meeting summaries. Studies demonstrate that compressing extractive summaries can raise ROUGE scores and human readability.



Keyword Extraction and Clustering for Document Recommendation in Conversations

Purity and Entrophy

Purity = 1 / N (w(j) * P(j))

Entrophy = 1 / N (w(j) * E(j))

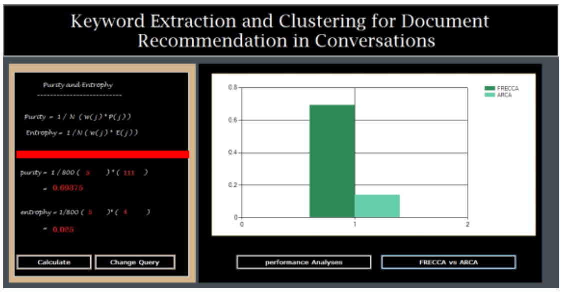
purity = 1 / 800 (5) * (111)

= 0.69375

entrophy = 1/800 (5)* (4)

= 0.025

Fig 2: FRECCA vs ARCA



Keyword Extraction and Clustering for Document Recommendation in Conversations

Purity and Entrophy

Purity = 1 / N (w(j) * P(j))

Entrophy = 1 / N (w(j) * E(j))

purity = 1 / 800 (5) * (111)

= 0.69375

entrophy = 1/800 (5)* (4)

= 0.025

Fig 3: Performance Analysis

5. CONCLUSION

We have provided a summary of automated keyphrase extraction's current state of the art. The challenge is far from being addressed, as seen by the relatively subpar state-of-the-art outcomes on different widely-used assessment datasets, even though unsupervised techniques have begun to match the performance of their supervised counterparts. According to our study, there are at least three significant obstacles to overcome. While most recent work has been on algorithmic advancement, in order for keyphrase extractors to function at the highest level, they need to have a deeper "knowledge" of a document. Background information can be included to aid in this comprehension. Even while it could be feasible to create algorithms that can manage the huge number of candidates in lengthy documents, we think that using advanced features—especially those that convey prior knowledge—will make it possible to discriminate between keyphrases and non-keyphrases more clearly even when there are many possibilities. Keyphrase extractors shouldn't be punished for assessment mistakes because this will allow us to more precisely gauge their performance. We have offered numerous solutions on how to deal with this issue.

DOI : https://doi.org/10.56452/7-2-562

REFERENCE

[1] M. Habibi and A. Popescu-Belis, "Enforcing topic diversity in a document recommender f or 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.Saltonand 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,andL.Qiu,"Document concept lattice for text understanding and summarization," Inf. Process. Manage. ,vol. 43,no.6,pp.1643–1662,2007.

[5] A.CsomaiandR.Mihalcea,"Linkingeducationalmaterialstoency-clopedic knowledge," in Proc. Conf. Artif. Intell. Educat.: Building Technol. Rich Learn. Contexts That Work ,2007,pp. 557–559.

[6]D.HarwathandT.J.Hazen,"Topici dentificationbasedextrinsiceval-uationofsummarizationtechniquesappliedtoconversationalspeech," 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 A MIDA automatic content linkingdevice:Just-in-timedocumentretrievalinmeetings,"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-timeretrievalsystemusingsemanticsearch,"in Proc. Annu. Conf. North Amer. Chap. ACL (HLT-NAACL),2011,pp. 80–85.

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

[10] B.RhodesandT.Starner,"RemembranceAgent:Acontinuouslyrun-ningautomatedinformationretrievalsystem,"inProc. 1st Int. Conf. Pract. Applicat. Intell. Agents Multi Agent Technol., London, U.K., 1996,pp.487–495.

[11] B.J.RhodesandP.Maes,"Just-in-timeinformationretrievalagents ,"
IBM Syst. J. ,vol.39,no.3.4,pp.685–704,2000.

[12] B. J. Rhodes, "The wearable Remembrance Agent: A system for
augmented memory," Personal Technol., vol. 1, no. 4, pp. 218–224,
1997.

[13] J.BudzikandK.J.Hammond,"Userinteractionswitheverydayap-plicationsascontextforjust-in-timeinformationaccess,"in Proc. 5th
Int. Conf. Intell. Us er Interfaces (IUI'00) ,2000,pp.44–51.

[14] M. Czerwinski, S. Dumais, G. Robertson, S. Dziadosz, S. Tiernan,
and M. Van Dantzich, "Visualizing implicit queries for information
management and retrieval," in Proc. SIGCHI Conf. Human Factors
Comput. Syst. (CHI),1999,pp.560–567.

[15] S.Dumais,E.Cutrell,R.Sarin,andE.Horvitz,"Implicitqueries(IQ)
forcontextuali zedsearch,"in Proc. 27th Annu. Int. ACM SIGIR Conf.
Res. Develop. Inf. Retrieval,2004,pp.594–594.

[16] M.Henzinger,B.-W.Chang,B.Milch,andS.Brin,"Query-freenews
search," World Wide Web: Internet Web Inf. Syst., vol. 8, no. 2, pp.
101–126,2005.