Trends in Extractive and Abstractive Techniques in Text Summarization

Neelima Bhatia
Amity School of Engineering and Technology (ASET)
Amity University Noida, India

Arunima Jaiswal
Amity School of Engineering and Technology (ASET)
Amity University Noida, India

ABSTRACT
Text Summarization was proved to be an advantage over manually summarizing the large data. It condenses the salient features from the text by preserving the content and serves the meaningful summary. Classification can be done in two ways – extractive and abstractive summarization. Extractive summarization uses statistical and linguistic features to determine the important features and fuse them into a shorter version. Whereas abstractive summarization understands the whole document and then generates the summary. In this paper extractive and abstractive methods are framed.

Keywords
Extractive summarization, Abstractive summarization

1. INTRODUCTION
The World Wide Web data is growing every second. It has become a tedious task to gather the information and then summarize it. Internet is the one such platform which retrieves the information from databases. But still this information is too large to handle. So text summarization came into demand which condenses the document into shorter version by preserving the meaning and the content.

A summary is thus helpful as it saves time as well as retrieves large documents data. Earlier humans read articles, documents and then understands and write their own summary. It can vary from person to person's understanding but consumes a lot of time. Therefore automatic summary served much better and is used in many fields like research, emails, news, messages, online information etc [1].

A good summary should be indicative as well as informative [2]. Indicative summaries points out some important parts while an informative emphasizes on the important information in a document.

The two main approaches used in summarization are extractive and abstractive. Extractive approach point outs the most important text from several documents and then fuse together and produce a summary. While abstractive approach understands the source text and outputs the precise and concise summary by using linguistic methods and compression techniques [3]. Earlier the summarization system was fed to only a single document. But with the bulk of information it moved to multi documents.

Evaluating a summary [9][10][11] is an important task for text summarization. Intrinsic measure evaluates the quality using human evaluation while extrinsic measure uses task-based [12] measure.

This paper presents the extractive and abstractive summarization approaches. The paper organization is as follows: Section 2 describes the problems in extractive and abstractive summaries. Section 3 describes the methods of extractive summaries. Section 4 describes the methods of abstractive summaries. Section 5 concludes the paper.

2. EXTRACTIVE and ABSTRACTIVE SUMMARIZATION
Extractive summaries do not focus on the understanding of text. It extracts the most important part based on statistical and linguistic features such as cue words, location, word frequency [4].

The process for extractive summarization could be [5]: 1) Pre-processing and 2) Processing.

The pre-processing phase is a structural framework of the text [6]. It consists of:

1) Sentence boundary identification:- identification of boundary is identified by the dot at the termination of a sentence.
2) Stop word elimination:- stop words and unnecessary information is discarded.
3) Stemming:- for every word a stem is build which gives meaning.

The processing phase calculates the relevant sentences and assigns the weights using weight learning method [6]. Then final scores are calculated by feature-weight equation and top ranked sentences are added in summary.

Extractive Summary Problems are as follows [7][8]:

- Sentences which are extracted are longer in length for summary and thus consume space.
- Not all the relevant sentences are included.
- Not much accurate information is presented.
- Coherency plays a big role as sentence or paragraph structure is disturbed while extracting. This is a severe problem in multi document extraction.

Abstractive Summary Problem is as follows [7]:

- Representation problem is the big issue

3. METHODS FOR EXTRACTIVE SUMMARIZATION
Extractive summaries [13][14][15] maintains the redundancy by extracting the relevant features from the document.

A. Term Frequency-Inverse Document Frequency method (TF-IDF):- Words are taken out with the help of weighted term-frequency and inverse sentence-frequency measure [16]. The similar

sentences are given scores and high scores are included in summary by information retrieval measure [6].

For generic summary, most frequently occurring non stop-words are taken as they represent the theme. Usually 0 or 1 is used for term-frequency [6].

B. Cluster based method:- Documents are written in an organized manner so that they can be divided implicitly as well as explicitly. This can be applied to summaries also. But if summaries follow some different themes, then clustering is required to give a proper meaning.

Document words are given scores using TF-IDF method [17]. IDF value is calculated over the document. The clustered document having TF-IDF score is fed to summarizer [6].

From cluster (Ci) as well as from location (Li) sentences are selected. Also the first sentence (Fi) in the document has important role in selection. Thus the total scores (Si) of a sentence i the weighted sum of (Fi, Ci, Li):-

where W1,W2,W3 are the combination weights.

C. Graph Theory approach:- Identification of themes is done [18]. The documents are represented as nodes when preprocessing, stop word removal and stemming is done. Every sentence is represented as a node. If two sentences share some common information then they are connected by edges. There could be sub-graphs that are not connected describes the distinct text of a document. For query-specific summaries sentences are selected from one sub-graph whereas for generic summaries, sentences are from different sub-graphs. Also if the node has a high cardinality number then it shows the importance.

D. Machine Learning approach:- Here sentences are classified on the basis of some features as summary and non-summary sentences. A training data is fed to the system [19] based on probabilities using Baye's.

E. LSA method:- SVD (singular value decomposition) [19] is called by different names like Karhunen Loeve transform, PCA (principal component analysis) and LSA (latent semantic analysis). It is a mathematical tool that works on multidimensional data. As SVD can be applied on document-word matrices, group documents so it is known as LSA. It easily captures the relations occurring in human brains.

F. Neural Network in text summarization:- Here neural networks are trained to learn the sentences. They learn the patterns so as to determine which sentence to be included in the summary or not [20]. It uses three-layered Feed forward neural network. When the neural network has learned the features, next step is to discover those features which are yet not taken by sentences [21]. This is done by feature-fusion phase:- 1) eliminate uncommon features and 2) fusing common features

G. Automatic text summarization based on fuzzy logic:- Various features like sentence length, similarity to title are fed to fuzzy system [4][22]. Then rules are entered in knowledge base of system. Then as the sentences are fed value from zero to one is obtained. This value is helpful in determining the important sentences. IF-THEN rules help to extract sentences. The input membership function has some insignificant values like low, very low, medium and significant values like high and very high. The performance is affected by fuzzy rules and membership function. There are four components in fuzzy system: fuzzifier, inference engine, defuzzifier and fuzzy knowledge base. In fuzzifier the membership function is used to translate inputs into linguistic values. Then IF-THEN rules are used to derive values. Then at last linguistic values are converted to final values.

H. Multi-document extractive summarization:- It deals with information which is extracted from multiple documents focusing on a single topic. Various opinions are taken from different users and are put together. This gives a concise and comprehensive result [6].

I. Query based extractive text summarization:- The sentences are given scores using frequency counts [23]. High scores are given to those sentences which contain query words and they are used in the summary. Also some part of summary is taken from different section or sub sections.

4. METHODS FOR EXTRACTIVE SUMMARIZATION

Abstractive methods are of two types: Structured based and Semantic based approach. Various methods like tree based, template based, ontology based, lead and body phrase and rule based use structured approach whereas multimodal semantic, information item based and semantic graph based use semantic approach.

1) Structure based approach:- It takes out the most important information by using cognitive schemas [24] like templates, extraction rules and various other structures like tree, ontology, lead and body phrase.

Tree based approach:- Dependency tree is used to represent text of a document [25]. Algorithm like theme intersection algorithm is used to select content for a summary. This either uses language generator or an algorithm to generate a summary [26]. Here shallow parser preprocess similar sentences and then they are mapped to predicate-argument structure. Then this is compared by using intersection algorithm. Thus some information is added to final phrases and is ordered. After that language generator gives a summary form by improving the quality.

In next approach, the dependency tree is obtained [27]. By finding the centroid basis tree is set and fuses it with sub-trees and the predefined constituents are pruned.

Template based method:- A template is used for a document containing slots and fillers. To identify text snippets linguistic patterns or extraction rules are matched.

These snippets are the indicators [25] and are extracted by Information Extraction Systems. It generates the well structured informative multi document summaries using multi document summarization algorithm. It works only with the information that is present in the document.

Ontology based method:- Ontology represents the domain which talks about the same topic having same knowledge. Here the domain ontology is defined by domain experts [28]. Next meaningful terms are produced by preprocessing and classifier classifies those terms. After that membership degree is generated this is linked with various events. But this process is only limited to Chinese data.

Lead and body phrase method:- This focuses on the phrases that has got same syntactic head chunk in lead and body sentences. Here the same chunks are searched in lead and body sentences [29]. Then these phrases are aligned using similarity metric. If the body phrase has rich information and has same corresponding phrase then substitution occurs. But if body phrase has no counterpart then insertion takes place. It has a drawback of rewriting the sentences.

Rule based method:- It is based on abstraction scheme [30]. To generate a sentence this scheme uses a rule based information extraction module, content selection heuristics and one or more patterns. To generate extraction rules similar meaning verbs and nouns are identified. Several candidate rules are selected and passed on to summary generation module. It provides the best summary but the main drawback is it consumes time as rules and patterns are written manually.

2) Semantic based approach:- Here the main focus is on identifying noun and verb phrases [31].

Multimodal semantic model:- This generates the abstractive summary from a semantic model [32]. The document is built up of text and images. Firstly by using knowledge a semantic model is build. Next the information is rated using information density metric which checks the completeness, relationship with others and number of occurrences of an expression. The expressions give the relationships and the concepts.

Information Item based method:- Here the content of summary is taken from the abstracts rather than from source information [25]. It gives the rich, unambiguous, structured and short summary.

Semantic Graph based method:- A semantic graph called Rich Semantic Graph (RSG) is build [33] where the nodes represents verb and nouns while edge give semantic and topological relationships. It uses heuristic rules to reduce the generated rich semantic graph to more reduced graph and thus abstractive summary is produced. It works on the semantics and gives less redundant and well structured summary.

5. CONCLUSION

In this paper all the extractive and abstractive summarization methods are reviewed with its pros and cons. An extractive summary deals with the important sentences while abstractive summary understands first and then builds the summary.

6. REFERENCES

[1] M. Haque*, et al.*, "Literature Review of Automatic Multiple Documents Text Summarization," *International Journal of Innovation and Applied Studies,* vol. 3, pp. 121-129, 2013.

[2] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg, 2005.

[3] D. Das and A. F. Martins, "A survey on automatic text summarization," *Literature Survey for the Language and Statistics II course at CMU,* vol. 4, pp. 192-195, 2007.

[4] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[5] Vishal Gupta, G.Sl Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009.

[6] Gupta V. and Lehal G.S. , “A Survey of Text Summarization Extractive Techniques”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[7] Jimmy Lin., “Summarization.”, Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009.

[8] Jackie CK Cheung, “Comparing Abstractive andExtractive Summarization of Evaluative Text: Controversiality and Content Selection”, B. Sc. (Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia, 2008.

[9] Ani Nenkova and ,Rebecca Passonneau, “Evaluating content selction in summarization: The Pyramid method”, in HLT-NAACL, 145-152, 2004.

[10] Chin-yew Lin, “A package for automatic evaluation of summaries”,in Proc. ACL workshop on text summarization branches out,2004.

[11] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto, “Automated Summarization Evaluation with Basic Elements”, In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006.

[12] Kathleen Mackeown, Ani Nenkova, David Elson, Rebecca Passonneau, and Julia Hirschberg “A task based evaluation of multidocument system”,in SIGIR’05, ACM, 2005.

[13] Madhavi K. Ganapathiraju, “Overview of summarization methods”, 11-742: Self-paced lab in Information Retrieval, November 26, 2002.

[14] Klaus Zechner, “A Literature Survey on Information Extraction and Text Summarization”, Computational Linguistics Program, Carnegie Mellon University, April 14, 1997

[15] Berry Michael W., "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43, 2004.

[16] Rene Arnulfo Garcia-Herandez and Yulia Ledeneva, "Word Sequence Models for Single Text Summarization", IEEE,44-48, 2009.

[17] Yongzheng, Nur and Evangelos, "Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora", WIDM'5, 51-57, Bremen Germany,2005.

[18] Canasai Kruengkari and Chuleerat Jaruskulchai, "Generic Text Summarization Using Local and Global Properties of Sentences", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03) , 2003.

[19] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach", Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.

[20] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", in Proceedings of second international Conference on intelligent systems, IEEE, 40-44, Texas, USA, June 2004.

[21] Khosrow Kaikhah "Text Summarization using Neural Networks", Department of Faculty Publications-Computer Science, Texas State University, eCommons,2004.

[22] Ladda Suanmali, Mohammed Salem, Binwahlan and Naomie Salim, "Sentence Features Fusion for Text summarization using Fuzzy Logic, IEEE, 142-145, 2009

[23] F. Canan Pembe and Tunga Güngör, "Automated Querybiased and Structure-preserving Text Summarization on Web Documents", Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, İstanbul, June 2007.

[24] P.E. Genest and G. Lapalme, "Framework for abstractive summarization using textto- text generation," in *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 2011, pp. 64-73.

[25] Khan A. and Salim N., "A REVIEW ON ABSTACTIVE SUMMARIZATION METHODS "Journal of Theoretical and Applied Information Technology 10th January 2014. Vol. 59 No.1

[26] R. Barzilay*, et al.*, "Information fusion in the context of multi-document summarization," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 550- 557.

[27] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," *Computational Linguistics,* vol. 31, pp. 297-328, 2005.

[28] C.-S. Lee*, et al.*, "A fuzzy ontology and its application to news summarization," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 35, pp. 859-880, 2005.

[29] H. Tanaka*, et al.*, "Syntax-driven sentence revision for broadcast news summarization," in *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, 2009, pp. 39-47.

[30] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers- Volume 2*, 2012, pp. 354-358.

[31] H. Saggion and G. Lapalme, "Generating indicative-informative summaries with sumUM," *Computational Linguistics,* vol. 28, pp. 497-526, 2002.

[32] C. F. Greenbacker, "Towards a framework for abstractive summarization of multimodal documents," *ACL HLT 2011,* p. 75, 2011.

[33] F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization," in *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*,2012, pp. 132-138.