IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 7, Issue 5, May 2018

A Survey of Automatic Text Summarization using Lexical Chains

Tapas Guha

Assistant Professor, Department of Computer Science, Presidency University, Bengaluru, India

Abstract: Text Summarization is a reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source. It is the process of distilling the most important information from a source (or sources) to produce an abridged version for particular user (or users) and task (or tasks). The process of producing summaries automatically is Automatic Text Summarization. This paper produces a survey of the cohesion based summarization technique: Lexical Chain.

Keywords: Automatic text summarization, Extracts, Cohesion, Lexical Chains.

I. INTRODUCTION

There are extensive application areas for automatic text summarization. With the information overload in the Internet it is getting increasingly difficult to navigate and select relevant information. Information is published in different versions across various media channels. A same news instance could, for instance, be published in an offline and online newspaper, a news channel, a SMS news flash, mobile radio newscast etc. Also, these may today be accessed by a myriad of display devices, sporting a wide range of presentation capacity. Customizing this information for different channels and formats is a monstrous editing job that obviously mandates the involvement of shortening of original texts. Automatic text summarization automates this work completely, or at least assist by producing a draft summary. Also, it can save human translators work by making documents accessible in other languages. It can first summarize documents before translation, which in many cases would be sufficient to establish the relevance of a foreign language document. Automatic text summarization can also reduce the time needed to absorb the key facts in a document, when used to summarize a text before an automatic speech synthesizer reads it.

Summaries can be of different types. It can be extractive (directly parts of the original text, produced verbatim) or abstractive (smaller number of concepts by fusing various concepts of the original text), indicative (keywords indicating topics) or informative (content laden), generic (author’s perspective, covering all important information of the document) vs. query-oriented (focus is driven by the user), single document or multi-document.

Much of the work to date has been in the context of generic summarization. Generic summarization makes few assumptions about the target audience or the goal for generating the summary. Typically, it is assumed that the audience is a general one: anyone may end up reading the summary. On the other hand, in query focused summarization, the goal is to summarize only the information in the input document(s) that is relevant to a specific user query.

In extractive summarization, different approaches have been evaluated recently. Extractive techniques merely copy those information from the text to the summary, which are deemed most important by the summarizer system e.g. word sequences like phrases, sentences or paragraphs and key clauses. Understandably, extractive summaries are mostly inconsistent, having lack of balance and output cohesion. Moreover, they may be out of context with broken anaphoric references.

II. APPROACHES

There are many text summarization approaches, some prominent ones are listed in the following table

TABLE 1
CLASSIFICATION OF TEXT SUMMARIZATION APPROACHES

Approach	**Example**
Statistical	Tf-Idf, Position of a Keyword etc.
Machine Learning	Maximum Entropy, HMM, Naïve Bayes, Neural Networks, Decision trees etc.
Coherent based	**Lexical chains**, Rhetorical Structure Theory (RST) etc.
Graph based	Hyperlinked Induced Topic Search (HITS), Google’s PageRank (GPR) etc.
Algebraic	Latent Semantic Analysis (LSA), Sentence level semantic analysis(SLSS), Non-Negative Matrix factorization (NMF and SNMF) etc

This paper focuses on the coherent based approach Lexical Chain.

DOI 10.17148/IJARCCE.2018.7522

II. PIONEERING WORKS

Current research in automatic text summarization largely views the process in two steps. The first step of the summarization process is to extract the important concepts from the source text into some form of intermediate representation. The second step is to use the intermediate representation to generate a coherent summary of the source document [1]. Till date, various methods have been proposed to extract the important concepts from a source text and to build the intermediate representation, as depicted in Table 1. Early methods focused primarily on word frequency to determine the most important document concepts, and thus were mostly statistical in nature [2]. The other extreme of such statistical approaches is to attempt genuine "semantic understanding" of the source text. Naturally the best chance to create a quality summary is the use of deep semantic analysis. The problem with such approaches is that a thorough semantic illustration has to be created and an area specific knowledge base must be available.

The major problem with purely statistical methods is that they never account for context. Specifically, finding the aboutness of a document depends largely on identifying and capturing the existence of not just duplicate terms, but related terms as well. This concept, known as cohesion, links semantically related terms which is an important component in a coherent text [3]. Lexical cohesion is the simplest form of cohesion. Lexical chains signify the lexical cohesion among a random number of related words. Lexical chains can be documented by recognizing sets of words that are semantically related (i.e. have a sense flow). Using lexical chains in text summarization is quite efficient, because these relations are easily distinguishable within the source text, and vast knowledge bases are not essential for computation. By using lexical chains, we can statistically find the most important concepts by looking at structure in the document rather than deep semantic meaning. All that is required to calculate these is a generic knowledge base that contains nouns, and their associations. These capture concept relations such as synonym, antonym, and hypernyms (is a relations).

III. LEXICAL CHAINS: LITERATURE SURVEY

A lot of attempts were made to understand the text instead of just extracting sentence based on relevance, not only structure. Quite naturally, capturing concept relations, anaphoric expressions etc. has been usually very difficult in extractive methods. In the beginning, *Roget's* International Thesaurus was used to manually construct the first lexical chains, by authors in [1]. They opined that given an electronic thesaurus, automation would be straightforward. [5] proposed the detection and correction of malapropisms using lexical chains. Another noteworthy attempt was made in by [6].The authors found out limitations in prior implementations of lexical chains. Potentially appropriate context information that follows a word is lostas all probable senses of the word are not considered, except at the time of insertion. The resulting problem is known as"greedy disambiguation". Authors in [6]presented a less greedy algorithm that builds all possible interpretations ofthe source document using lexical chains. Their algorithm then selects the interpretation with the strongest cohesion. These "strong chains" are then utilized to produce a summary of the original text. Using WordNet they produced Lexical chains which provide a representation of the lexical cohesive structure (relations like repetition, synonymy, antonymy and holonomy) of the text. Their scores were determined based on the number as well as type of relations in the chain. Only those sentences with the highest concentration of the strongest chains were selected for the summary.

[8] presented an algorithm for calculating lexical chains in linear time. The algorithm presented is clearly O(n) in the number of nouns present within the source document. Including generation, a 40,000 word corpus was summarized in eleven seconds, in tests conducted on a Sun Sparc Ultra10 Creator.
Authors in [9]proposed an approach to find word sense disambiguation using two relations: same word repetition and same head word inclusion, using Roget's Thesaurus. Then the word is inserted in the chain via thesaurus relation. In [10] authors used lexical chains to summarize Chinese texts, based on the HowNet knowledge database. Moreover, the construction rules of lexical chains are extended, and relationship among more lexical items is used. The algorithm constructs lexical chains first, and then strong chains are identified and significant sentences are extracted from the text to generate the summary. Evaluation results show that the performance of the system has a notable improvement both in precision and recall compared to the original system.
[11] proposed an automatic system for extracting key points from documents by using lexical chains using FrameNet for shallow semantic parsing of texts. The chain is scored using four distinct features. In 42 percent of the cases, the concept which generated by this system is equal to the concept generated by human.
In [12], authorspropose a new algorithm for Lexical Chaining, based on a global function optimization through Relaxation Labelling. A preliminary evaluation of the performance of our approach has been performed on a Catalan agency news corpus. The results in a preliminary evaluation show that the presented approach outperforms other algorithms in the score of the found chains, with only a minor increase in runtime. The resulting lexical chain has been used for a complete multilingual Automatic Summarization system, available on-line.

DOI 10.17148/IJARCCE.2018.7522

Using WordNet lexical corpus, [13]proposed lexical chain analysis using linguistic pre-processing. This includes sentence segmentation, tokenization, POS tagging, entity detection, relation detection respectively in order. They generated lexical chain using the candidate sets they extracted in the form of nouns and noun compounds. Their proposed algorithm of lexical chain generation is as follows:

```
For each word in the candidate set {
For each chain in lexical chain {
Find Word sense between two words
}
If (distance > Threshold)
Add word to that chain
Else
Generate new chain.
}
```

In [14], the authorsproposed an approach which does not require full semantic interpretation of the text, instead creates a summary using a model of topic progression in the text derived from lexical chains using WordNet thesaurus. Further, they also overcome the limitations of the lexical chain approach to generate a good summary by implementing pronoun resolution and by suggesting new scoring techniques to leverage the structure of news articles.
The following table gives a comparative snapshot of the different prominent text summarization approaches using lexical chains, chronologically.

TABLE 2
COMPARISON OF SUMMARIZATION APPROACHES USING LEXICAL CHAIN

Authors	Algorithm	Year	Candidate set	Word Sense Disambiguation
Morris and Hirst	Lexical Chain from corpus	1991	All words, repetitions	Roget’s Thesaurus.
Hirst and St-Onge	Lexical Chain from corpus, correction of malapropisms	1995	All words except stop-words,	WordNet as a knowledge source.
Brazilay and Elhadad	Dynamic chaining	1997	All nouns and noun compounds	WordNet, Systematic relations
Kevin Humphreys and Robert Galzauskas	Coreference chains	1999	Set of nouns with co-reference chains	‘best parse’ of two sentences
Silber, H., McCoy, K	Linear Lexical chain	2002	All nouns	Synonyms, hyponyms and hypernyms from WordNet
Mario Jarmasz, Szpakowicz	Head Words Lexical chain	2003	Head words	Same word repetition and head word from Roget’s Thesaurus
Yanmin Chen, Xiaolong Wang, and Yi Guan	Multilevel lexical chains	2005	--	HowNet word database
Mohamadi, Sudabeh & Badie, Kambiz & Moeini, Ali	Frame based Lexical Chains	2011	--	FrameNet
Edgar Gonzàlez & Maria Fuentes	Global lexical chains with relaxation labelling	2014	--	Wordnet
Patel, Dabhi & Prajapati	Linguistically pre-processed lexical chain	2017	Nouns and noun compounds	WordNet
Sethi, Sameer et al		2017	All nouns and Pronoun resolution	WordNet

V. CONCLUSION

This paper presents a detailed review of the popular text summarization method Lexical Chain. Chronologically this work surveys the variations of the technique, the algorithm, the candidate seta and the word sense disambiguation employees by the respective authors in that particular algorithm.

ACKNOWLEDGMENT

The author is grateful to the Presidency University, Bengaluru for providing the essential facilities and infrastructure for the conduct of the study. The author would like to thank the proof-readers and the anonymous reviewer for their constructive comments and suggestions in improving the quality of this paper.

REFERENCES

[1] Jones, Karen Sparck,"What might be in summary?" Information Retrieval, 1993.
[2] Luhn, H.P., "The automatic creation of literature abstracts," In H.P. Luhn: Pioneer of Information Science. Schultz, editor. Spartan, 1968.
[3] Halliday, Michael and Ruqaiya Hasan. Cohesion in English. Longman, London, 1976.
[4] Morris, J. and G. Hirst, "Lexical cohesion computed by thesaurus relations as an indicator of the structure of the text," In Computational Linguistics, 18(1):pp21-45. 1991.
[5] G. Hirst, D. St-Onge, "Lexical chains as representation of context for the detection and correction of malapropisms," in C. Fellbaum (Ed), WordNet: An electronic lexical database and some of its applications, The MIT Press, Cambridge, MA, 1995.
[6] Barzilay, Regina and Michael Elhadad. "Using Lexical Chains for Text Summarization," in Proc. The Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL Madrid, 1997.
[7] Saliha Azzam , Kevin Humphreys , Robert Gaizauskas, "Using coreference chains for text summarization," in Proc. The Workshop on Coreference and its Applications, College Park, Maryland, June 22-22, 1999
[8] H.G. Silber, K.F. McKoy, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," Computational Linguistics, 28(4), pp. 487–496, 2002.
[9] Mario Jarmasz and Stan Szpakowicz, "Not as easy as it seems: Automating the construction of lexical chains using rogets thesaurus" Advances in Artificial Intelligence, 2671(2003):994–999, 2003
[10] Chen Y., Wang X., Guan Y, "Automatic Text Summarization Based on Lexical Chains. In: Wang L., Chen K., Ong Y.S. (eds)" Advances in Natural Computation. ICNC 2005. Lecture Notes in Computer Science, vol 3610. Springer, Berlin, Heidelberg, 2005
[11] Mohamadi, Sudabeh & Badie, Kambiz & Moeini, Ali, "Using Frame-based Lexical Chains for Extracting Key Points from Texts" in ProcThe Third International Conference on Creative Content Technologies, 2011
[12] Gonzàlez, E., Fuentes, M, "A New Lexical Chain Algorithm Used for Automatic Summarization" In: Proc. The 12th International Congress of the Catalan Association of Artificial Intelligence (CCIA), 2009
[13] Patel, Sagar M., Vipul K. Dabhi and Harshad B. Prajapati. "Extractive Based Automatic Text Summarization." *JCP* 12 : 550-563, 2017
Sethi, Sameer et al, "Automatic text summarization of news articles," in Proc. IEEE International conference 2017.

BIOGRAPHY

Tapas Guha is working as an Assistant Professor in the Dept. of CSE, Presidency University, Bengaluru, India. He is M.Tech from IIT Kharagpur and currently pursuing his Ph.D. in Data Science under the guidance of Dr. N Mehla, Associate Professor, Presidency University. His research interest includes Social Media Analysis, opinion mining and sentiment analysis. He has around 13 years of academic experience along with 3 years of industry experience.

DOI 10.17148/IJARCCE.2018.7522