RESEARCH ARTICLE OPEN ACCESS

A survey on sentence fusion techniques of abstractive text summarization

R.V.V Muralikrishna, Dr. Ch Satyananda Reddy

Information Technology, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, Andhra Pradesh, India.
Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.

Abstract

Sentence fusion is one of the tasks of automatic text summarization that has wide spread applications in the field of computer science. It is currently used in automatic text summarization and question answering systems. It is also used in natural language generation systems in the name of sentence aggregation. The task of sentence fusion poses various challenges like deciding whether two sentences can be combined or not, which part of a sentence should be selected for combination, how these parts can be combined and how to fit the combinations in a grammatically correct sentence structure. In this paper we discuss about the research done on sentence fusion and various challenges that are yet to be met.

Index Terms– Automatic text summarization, Abstractive text summarization, Natural language processing, Sentence fusion.

I. INTRODUCTION

Automatic text summarization is the science of creating a concise and meaningful text to represent a given text. It can be broadly classified in to two types, Extractive text summarization and Abstractive text summarization. Extractive summarization produces Extractive summary, which is a collection of key sentences of the original text. Abstractive text summarization produces abstractive summary, which is a collection of sentences not present in the original text but represent the meaning of the original text.

Sentence compression and sentence fusion techniques have been suggested in the literature for abstractive text summarization. Sentence compression [1] involves strategies like replacing a sentence or phrase in the original text with a one word substitute and deleting unnecessary words in a sentence or phrase in the original text. The term sentence fusion is first introduced by Regina Barzilay et al. [2] in the year 2005. They defined sentence fusion as a novel text-to-text generation technique for synthesizing common information across documents. In simpler words Sentence fusion is the technique that converts the information from two or more sentences in to a single output sentence. The concept of Sentence fusion has existed in the past also but with different names. It was earlier used for automatic text summarization by Regina Barzilay et al. [3] in the name of *information fusion.* Hongyan Jing et al. [4] have also used sentence fusion for automatic text summarization and they referred sentence fusion as *sentence combination*.

Sentence fusion can be categorized into two types, Union fusion and intersection fusion, based on the information in the generated output sentence [5]. In the union fusion, output sentence contains the information present in both the sentences. In intersection fusion the output sentence contains the common information present in both sentences.

The sentence fusion method proposed by Regina Barzilay et al. [2] is an example of intersection fusion. Erwin Marsi et al. [5] have developed a methodology to use information in two or more sentences to generate a new sentence.

Examples for Intersection fusion and union fusion [6]:

Consider two sentences

- Posttraumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.
- Posttraumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psycho trauma) that can develop after exposure to a terrifying event.

1 ***Intersection fusion****:* Posttraumatic stress disorder (PTSD) is a psychological disorder.

2 ***Union fusion:*** Posttraumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psycho trauma) that can develop after exposure to a terrifying event.

Hal Daum´e III et al. [7] did not accept generic sentence fusion as a well defined summarization task. In their experiment a group of persons are assigned the task of fusing consecutive sentences extracted from a document. Each of these persons gave a different output sentence. All these persons have employed a different methodology for fusion and their priority of the topics in the input sentences is different. From this experiment Hal Daum´e III et al. concluded that generic sentence fusion is an Ill-Defined Summarization Task. Emiel Krahmer et al. [6] have shown that query-based sentence fusion is a better defined task than generic sentence fusion. While using Query-based fusions of various pairs of sentences they got short and meaningful output sentences. In their experiment with query-based sentence fusion, both intersection and union fusion strategies were considered. The results were compared with generic sentence fusion.

Below are some of the results obtained from their experiment (sentences 1 and 2 described above are inputs to the experiment).

3. ***Generic Intersection:***

Result: Posttraumatic stress disorder (PTSD) is a psychological disorder.

4. ***Query-based Intersection:***

Input query: What is PTSD?

Result: PTSD stands for posttraumatic stress disorder and is a psychological disorder.

5. ***Generic Union:***

Result: Posttraumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psycho trauma) that can develop after exposure to a terrifying event.

6. ***Query-based Union:***

Input query: What is PTSD?

Result: PTSD (posttraumatic stress disorder) is a psychological disorder caused by a mental trauma (also called psycho trauma) that can develop after exposure to a terrifying event.

A. Dependency trees

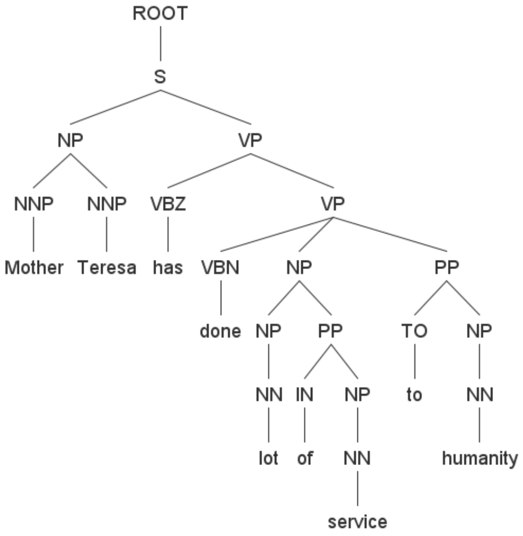
The processing of natural language text involves Parsing of sentences in the text. Parsing is done to identify the Parts of Speech of each word in the sentence and the grammatical relations that exist in between the words of the sentence. The output of parsing is usually a tree that gives a visual picture of grammatical relations and dependencies between the words in the sentence. These trees are called as dependency trees.

Dependency trees are used in many natural language processing applications. They play a prominent role in sentence fusion. Sentences to be fused are first represented as dependency trees.

These trees are merged in to a single tree and finally the merged dependency tree is converted to a sentence which is known as the fused sentence. The process of converting a dependency tree in to a string of words is called as tree linearization.

There are two prominent approaches to tree linearization [8]. First method is based on picking up a best sentence from all possible sentences that can be generated from the tree. Second method relies on the output of a machine that considers language modeling techniques and linguistic features of the language to generate an appropriate sentence.

Example:
Sentence: Mother Teresa has done lot of service to humanity.



II. DEVELOPMENTS IN SENTENCE FUSION

Before sentences are fused it should be determined whether they can be fused or not. A fused sentence can be generated only with conceptually coherent input sentences. So selection of sentences plays a key role. Hongyan Jing et al. [4] focused on combining manually selected key sentences from the original text. Much attention has been paid to the tasks like ‘When to use sentence combination’ and ‘which operator to use for combining sentences’. Sentence fusion is more complex than sentence combination as it aims for not only combining the sentences but also for compressing the overall data.

A. Sentence fusion for Multi document summarization

Regina Barzilay et al. [2] has developed sentence fusion methodology in the context of multi document summarization and is based on the below mentioned steps.

Step 1: Identification of common Information.
Step 2: Fusion lattice computation.
Step 3: Lattice linearization.

1) Identification of Common Information

In this step concepts shared by both the sentences are identified. This is done by representing the sentences by dependency trees [9] and then matching the parse trees. Dependency trees are well designed with information of each word in the sentence stored at each node. For example a node contains the parts of speech information of a particular word and its relation to other words. The alignment of sub trees is dependent on similarity between the structure of the dependency trees and the similarity between lexical items. The structural similarity takes in to account the type of edges between the nodes of the tree. Type of the edge is based on relationship between the nodes like subject-verb. Semantic similarity is based on WordNet [10] [11] and paraphrasing dictionary [3].

2) Fusion Lattice Computation

In this step the aligned sub trees are combined in to a single sub tree. Techniques like Paraphrasing, removing / adding phrases are used in creating the single sub tree.

3) Lattice linearization

Lattice Linearization is the final task in sentence fusion. This step uses the parse tree generated in previous phase as the basis for sentence generation. Some of the challenging tasks are selection of a tree traversal order, lexical choice among available alternatives, and placement of auxiliaries, such as determiners.

B. Sentence fusion for Question and answering systems

Erwin Marsi et al. [5] worked on sentence fusion in the context of Question and answering systems. They used fusion lattice computation and lattice linearization to generate a fused sentence. Entailment and any partial content overlap are also considered in fusion lattice computation stage in addition to structural similarity and semantic similarity. In the lattice linearization stage the type of fusion (union or intersection) is determined before the sub trees are merged.

C. Sentence fusion based on Graph Analysis

Katja Filippova et al. [12] used integer linear programming to improvise dependency trees to Directed Acyclic Graphs (DAG) which have syntactic importance and word information. The co arguments in the resulting sentence are checked for compatibility in terms of syntax and semantics. Optimal tree is selected from the Directed Acyclic Graph based on well-formedness of the sentence. The overall process is unsupervised.

D. Sentence fusion based on dependency trees

Micha Elsner et al. [13] used dependency trees for representing sentences. The task of aligning the sub trees and task of merging them are unified in to a single supervised optimization problem. The system is trained by the sentence fusion examples given by professional journalists.

III. IMPORTANT ISSUES IN SENTENCE FUSION

A. Selection of sentences for sentence fusion

Deciding whether two or more sentences can be combined or not is an important task and the decision depends on various parameters like Context, tense, multiple opinions, keywords/ theme computation.

Context: To combine two sentences the words in the sentences should have the same context. Because words convey different meanings when they are used in two different contexts

Tense: Sentences should be carefully fused if they are of different tenses. Here the problem occurs in selecting a tense for the fused sentence.

Date: Sentences involving dates have to be carefully combined. If the sentences are not sequential with respect to time, the meaning conveyed by the fused sentence is not coherent with respect to the original text.

Opinions: Sentences expressing the opinions of two persons on the same subject can be combined only if they have similar bias. Conflicting opinions cannot be fused.

B. Selection of mode of sentence fusion

According to Erwin Marsi et al. [5] one of the challenges in sentence fusion is to decide in between union fusion and intersection fusion. Based on this decision the contents of the output sentence are dependent.

C. Sentence generation

In English same words can form entirely different sentences when the combination is altered. Given keywords/ theme generation of a sentence is a challenging task.

D. Evaluating the output sentence

There are six basic qualities to be considered in evaluating output sentence of a sentence fusion operation [12]. They are

1) Meaning**:** Original meaning in the source sentences should be retained in the output sentence.
2) Clarity: There should be clarity in the expression of the concept.
3) Coherence: All parts of the sentence should be coherent with each other i.e. one part should not contradict other.

4) Emphasis: The key words and phrases should be placed in proper positions.
5) Conciseness: The output sentence should contain minimum number of words.
6) Rhythm: Flow of concept in the sentence is important. If it is missing it is difficult to understand the meaning of the sentence.

E. *Theme computation*

There are various sentence scoring methods developed to find similarity between two sentences. Statistical methods find similarity using common words in both the sentences. Semantic methods consider the meaning of words in both the sentences. Statistical and Semantical similarity between words do not exactly represent the theme of the sentences.

F. *Dataset*

Kathleen McKeown et al. developed a dataset to evaluate automatic sentence fusion [14]. The dataset consists of 297 pairs of sentences. For each pair of sentences in the dataset, fused sentences are created manually by different groups of professionals separately. Fused sentences are generated with intersection fusion and union fusion methodologies. When the pairs of input sentences in the dataset are observed , it is found that 237 pairs of sentences have at least 50% of words in common. So it can be said that statistical similarity can play a role in checking the validity of sentences before fusion.

IV. CONCLUSION

Sentence fusion is a higher order cognitive operation as its success much depends on how the sentences are understood. Sentence parsing techniques like dependency tree provide some degree of understanding of the sentences. Proper alignment of dependency tree is crucial for determining the portions of sentences that enable fusion. As the fusion operation is performed on the dependency trees the result will also be a tree which should be converted to sentence format according to the rules of the grammar. Sentence fusion can be best utilized when appropriate sentence fusion methodology (i.e, Union or Intersection fusion) is chosen. Query based sentence fusion is useful for building automatic question answering systems, abstractive summarization systems.

REFERENCES

[1] Kiwamu Yamagata, Satoshi Fukutomi, Kazuyuki Takagi, , Kazuhiko Ozeki , "*Sentence Compression Using Statistical Information About Dependency Path Length*", Text, Speech and Dialogue., Springer, Volume 4188, 2006, pp 127-134

[2] Regina Barzilay and Kathleen McKeown. Sentence Fusion for Multidocument News Summarization. Computational Linguistics, Journal Computational Linguistics archive Volume 31 Issue 3, September 2005, pp 297-328.

[3] Regina Barzilay, K. McKeown, and M. Elhaded. Information fusion in the context of multidocument summarization. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), Maryland, 1999, pp 550-557.

[4] Hongyan Jing and Kathleen R. McKeown. Cut and Paste Based Text Summarization. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, 2000, pp178-185.

[5] Erwin Marsi and Emiel Krahmer. Explorations in sentence fusion. In Proceedings of the 10th European Workshop on Natural Language Generation, 2005, pp 109–117.

[6] Emiel Krahmer, Erwin Marsi, Paul van Pelt. Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. In Proceeding HLT-Short '08 Proceedings of the 46th Annual Meeting of the Association for

Computational Linguistics on Human Language Technologies, 2008, Short Papers Pages 193-196.

[7] Daum'e III, Hal and Marcu, Daniel. Generic Sentence Fusion is an Ill-Defined Summarization Task. In proceedings of ACL Summarization, Association for Computational Linguistics, July 2004, pp 96-103.

[8] Katja Filippova and Michael Strube, Tree linearization in English: Improving language model based approaches. In Proc. NAACL HLT, 2009, pp 225-228

[9] Mel'cuk I. Dependency syntax: theory and practice. New York: SUNY University Press. 1988.

[10] Christiane Fellbaum. "*WordNet: An Electronic Lexical Database. Cambridge*", MA: MIT Press, 1998.

[11] George A. Miller. " *WordNet: A Lexical Database for English*". Communications of the ACM Vol. 38, 1995, No. 11: pp.39-41.

[12] Katja Filippova and Michael Strube. Sentence Fusion via Dependency Graph Compression. In Proceedings of the Conference on Empirical Methods in Natural Language Processing Honolulu, October 2008, pp 177–185.

[13] Micha Elsner and Deepak Santhanam. Learning to Fuse Disparate Sentences. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 24 June 2011, pp 54–63.

[14] http://www.cs.columbia.edu/_kathy/fusion corpus