Summarizing with Wikipedia

Abdullah Bawakid and Mourad Oussalah

School of Engineering
Department of Electronic, Electrical and Computer Engineering
University of Birmingham
{ axb517 , M.oussalah }@bham.ac.uk

Abstract:

This paper describes a query-based multi-document summarizer that was built to participate in the update summarization task of TAC10. The system relies on a thesaurus extracted from Wikipedia and uses it as its underlying ontology. The concepts which are detected within the documents are used as weighted features to score the document sentences. The relationships previously defined in the thesaurus between the different concepts help in finding the most important concepts within a document or a set of documents. Sentences are ranked based on the scores they have been assigned and the summary is formed from the highest ranking sentences till the 100-word limit is reached. The evaluation results and the performance of the system are described. The system's rank is the 7th in the manual evaluation of the update task for this year. The total number of the submitted runs by all participants is 43.

1 Introduction

The Text Analysis Conference (TAC) is one of the well-known workshops in the field of Natural Language Processing which provides the infrastructure necessary to evaluate different methodologies with different tasks. In TAC10, we participated in the Guided Summarization task with two different runs. The aim of the task is to provide short summaries for a set of newswire articles. The generated summaries are not to exceed 100 words each. This year's task is different from last year in that the participants are asked to a deeper semantic analysis of the source documents instead of simply relying documents words frequencies to select the important concepts. For this, a list of categories and important aspects for each category are given and it is asked that the summary provided should cover all of the mentioned aspects if possible in addition to any other information related to the topic.

The "update" part of the task is similar to that of TAC09 and TAC08. For a given set of documents, the participants are asked to write two summaries, one for set A and another for set B. A topic statement is provided in addition to the Categories aspects which have been added to the task only this year. The participants are asked to write 100-word summary for set A using the given topic statement and the specified category. For set B, a 100-word update summary is to be generated assuming that the user has already read the set of articles in set A.

To enhance the representation of the documents to summarize in each set, the developed system described in this paper applies a set of rules to expand the document representation with the help of an external ontology. In our participation in the Summarization task of TAC08, we relied on WordNet as an external ontology [1]. In this year, we used Wikipedia instead. Wikipedia has several advantages over what WordNet has to offer. The coverage and breadth of Wikipedia is larger than that of WordNet. In addition, it is more up-to-date. Using the concepts extracted from Wikipedia is especially useful with

short topic statements provided in the task for each set. Also, they are used to detect the most dominant concepts within a document and the inter-connection between these dominant concepts within a document set and the given topic.

In our system, we use the Wikipedia ontology to build a thesaurus containing a list of Wikipedia's concepts. To determine the relationship between all of the extracted concepts from Wikipedia, we used the internal links, categories structure and other rules. These concepts are used to aid in extracting the dominant concepts within each document and documents set, and the association strength of the extracted concepts. The ontology we used along with a description of how it was built was reported in our earlier work[2].

The rest of this paper is structured as follows: an overview of the related work followed by a description of how the concepts ontology was built and extracted from Wikipedia. Then, we describe our system and how it was applied to this year's task. Next, we present the evaluation results and discusses the rank, the strength and the limitations of our system. Finally, the paper is concluded with a potential future work.

2 System Overview

The system developed for the summarization task is extractive. Each sentence is assigned a score signifying its importance based on its extracted features. The summary is then generated for sets A by ranking the sentences based on their assigned scores in a descending order and choosing the top n sentences till the maximum word-limit is reached. The stages involved for creating summaries are summarized in the following subsections:

2.1 Preprocessing

The first stage in the framework is to preprocess all fed documents by cleaning them and then parsing them to extract the text and topics and then tokenizing the terms and splitting the sentences. The stop words are then removed.

3.2 Identifying the concepts

Two methods have been utilized to detect concepts by employing the built Wikipedia-thesaurus and its extracted features. First one is through an exact match measure where explicitly mentioned concepts within each sentence are detected. A concept having multiple spellings for a concept and synonyms should still be detected by the system as a single concept. This is due to the integration of redirect links within the thesaurus and the mapping algorithm that associates sentences with the concepts they contain. As for ambiguous terms and concepts, the system implements the Weighted Strong Links method that was described in [2].
In the second method we examine each term within a sentence and replace it with its concepts vector is through the term-concepts table. The concepts vector has a weight associated to each concept signifying its relatedness with the term. After generating a concepts vector for each term, we group all concepts vectors within a sentence by summing the scores of the individual concepts that are repeated. This in effect applies word sense disambiguation as relevant concepts are boosted and given a higher score in the merged concepts vector. For example, the concept "Fox" has two meanings: "Fox (Animal)" and

"Fox (Broadcasting company)". Similarly, the concept "Dog" is associated with "Mammals". In the sentence "A fox attacked a dog", the meaning "Fox (Animal)" is boosted.

3.3 Feature Selection

Each sentence is tagged with several features. These features are used to compute a score determining the sentence importance.

Overlap with the Topic: In our system, we consider the overlap between each sentence and the topic of its document set. We take into account both the concepts overlap and the terms overlap when assigning a score to each sentence. Synonyms and concepts with alternative spellings are considered as a single concept in our system with the help of the Wikipedia thesaurus and the custom matcher.

Concepts Dominance: The explicitly mentioned concepts within a document set which are most frequent and the topic concepts are considered to be the most important. When computing a score for each sentence based on this feature, we consider how pertinent the sentence concepts to the important concepts with the document set. We use the relevancy degrees between the concepts which are precomputed in the Wikipedia thesaurus for achieving this task.

Sentence Position: The system assumes that sentences appearing at the top and bottom of a document have more chances of being important than the rest. Therefore, sentences appearing in the top 20% and the bottom 20% portion of a document are given position scores 50% larger than the others.

3.4 Measuring the Relatedness and Similarity between Sentences

Each sentence would have a vector of the concepts detected in it using the exact match method. In addition, it would have another vector of concepts generated from merging its individual terms concepts as extracted from the term-concepts table. When evaluating two sentences, we consider both vectors to compute the similarity and relatedness between them. The semantic relatedness is computed by the following formula:

$$Srel(Sent1, Sent2) = \frac{rel(A, B)}{PairsCounter}$$

Where Sent1 and Sent2 refer to Sentence1 and Sentence2 respectively, A is the concepts set in Sentence1, B is the concepts set in Sentence2, and PairsCounter is the number of concepts pairs compared. This formula can be applied to both vectors individual.

3.5 Summary Generation

Knowing what features to use in the system, it is possible to assign a score for each feature in each sentence. A sentence score comprises of its Topics scores, the relevancy of these Topics with the dominant ones, the overlap between the sentence and the rest of the sentences in a document, and the position of a sentence in the document. After scoring all sentences, the summary is formed by ranking the sentences in a descending order based on their scores, and adding the sentences one by one to the summary till the 100-word limit is reached.

After adding the last sentence to the summary and reaching the mentioned word limit, the sentences are re-ordered according to their appearance in the original documents they were taken from. The last sentence in the summary is then truncated to enforce the 100-word limit. At last, we applied a custom set of rules we developed to remove non-important data from some sentences such as date stamps and writers references appearing at the beginning of some sentences.

4 Evaluations

The provided dataset for the update task is composed of 46 topics divided into five categories. Each topic has a title, category, and 20 relevant documents divided equally into two sets: A and B. Documents in set A precede chronologically those in set B. Participants are asked to submit a summary for each set. They are also given the option of submitting up to two runs for each team.

We participated with two runs. The ids of our runs are 14 and 19. The term-concepts table method was used with run 19 while strong links method was used for 14. In Table 1, the ranks obtained by the system in the different evaluation methods are displayed. The total number of runs the system is compared with is 43.

	Manual	ROUGE-2	ROUGE-SU4	BE
Run 14-A	7	18	14	12
Run 19-A	7	19	17	15
Run 14-B	8	16	14	15
Run-19-B	10	10	16	18

Table 1: Evaluation results for the Update Task showing ranks of the two submitted runs 14 and 19 relative to the 43 submitted runs

5 Conclusion

In this paper, we briefly described the methodology that was implemented in our system for this year's Update task. We outlined how Wikipedia was used, the features that we focused on, and how the summaries were constructed. The results obtained show that the performance of our system is competitive when compared with the other teams systems, although there is still room for improvement. Creating a redundancy/diversity matcher and finding a better method to set their thresholds, and implementing better measures to utilize the found concepts and better understand what they refer to through a deeper linguistic analysis than what is performed here are potential future work we intend to focus on.

References:

[1] A. Bawakid and M. Oussalah, "A Semantic Summarization System: University of Birmingham at TAC 2008," in *Proceedings of the First Text Analysis Conference (TAC 2008)*, 2008.

[2] A. Bawakid and M. Oussalah, "Centroid-based Classification Enhanced with Wikipedia," in *The Ninth International Conference on Machine Learning and Applications 2010*, 2010.

[3] L. Qiu, M. Kan, and T. Chua, "A Public Reference Implementation of the RAP Anaphora Resolution Algorithm," *cs/0406031*, Jun. 2004.

[4] A. Bawakid and M. Oussalah, "Using Features Extracted from Wikipedia for the Task of Word Sense Disambiguation," in *9th Conference on Cybernetic Intelligent Systems*, 2010.