Chapter 5

Applications of Triplets and Semantic Networks

This chapter will discuss the possible applications of triplets and semantic networks extracted from text in different contexts. First we present an application of multi-document summarization that we developed making use of the triplets extracted by our system pipeline discussed in Chapter 2 and also evaluate the application using standard metrics. In the following sections we discuss about other application areas where triplets and semantic networks could be useful, providing references to related work in literature.

5.1 Multi-Document Summarization

With the continuing growth of online information, it has become increasingly important to provide improved mechanisms to find and present textual information effectively. Over the past several years, there has been much interest in the task of multi-document summarization. Multi-document summarization aims to present multiple documents in form of a short summary. This short summary can be used as a replacement for the original documents to reduce, for instance, the time a reader would spend if she were to read the original documents.

Following the trends in summarization research, we focus on extractive summarization which simplifies the summarization task to the problem of identifying

key sentences from a set of documents which are then concatenated to create the summary. Nenkova and McKeown (2012) have conducted an extensive survey of text summarization techniques. According to them, summarization systems perform three steps for extractive summarization of a given text: 1) Using an intermediate representation for the text which captures its key features, 2) Using the intermediate representation to assign scores for individual sentences within the text and 3) Selecting a set of sentences which maximizes the total score as the summary for the targeted text. Event-based summarization is of great interest in recent research. Filatova and Hatzivassiloglou (2004) and Li et al. (2006) show that atomic events, which are the relationships between the important named entities can be automatically extracted from text and used for summarization, while describing algorithms that utilize this feature to select sentences for the summary.

Liu et al. (2009) discusses a method for extractive summarization using sentence extraction, grouping of sentences with similar content using hierarchical clustering, scoring of sentences based on features such as sentence length, number of keywords in a sentence and sentence compression to condense the summary. Darling (2010) discusses a summarization system called SumBasic+ which extracts sentences which contains the highest frequency words. Sentences are scored based on a function with the unigram probability distribution for the words in that sentence and iteratively the highest scoring sentences are included in the summary. Instead of using word frequencies the work by Conroy et al. (2006) directly models the set of terms/vocabulary that is likely to occur in a sample of human summaries using query terms (terms extracted from topic description) and signature terms (terms extracted from documents). Another work by Gong et al. (2009) uses Wikipedia to extract three concept features for each sentence along with sentence position and sentence length features and use them to generate a sentence based extractive summary using the MEAD (multidocument multilingual text summarization) (Radev et al., 2004) platform.

Rusu et al. (2009) extracts subject, verb, object triplets out of text, creates a semantic network out of it and use this graph to construct document summaries. Triplets are assigned a set of features consisting linguistic, document and graph attributes and a linear SVM classifier is trained in order to select

the triplets that are useful for extracting sentences for the summary. The linguistic attributes include, the type (subject, verb or object), the depth of the linguistic node extracted from the Treebank parse tree (treebank parsers include the Stanford Parser (Klein and Manning, 2003a) and OpenNLP) and the part of speech tag. Document attributes include the location of the sentence within the document, the triplet location within the sentence, the frequency of the triplet element, the number of named entities in the sentence and the similarity of the sentence with the centroid (the central words of the document). Graph attributes consist of hub and authority weights, page rank, node degree etc. This system ranked 17 out of the 25 systems in DUC 2007.

We draw insights from this approach and use the triplets extracted by our system pipeline discussed in Chapter 2 to generate automatic summaries. The method combines sentence scoring based on the identification of key entities and actions to rank sentences, and choose the ones that are most appropriate to go into the summary. We only use a few linguistic features to extract senteces for the summary without the training of a linear classifier. The details of the method are discussed in section 5.1.3.

5.1.1 Dataset

The dataset used for this task was from the guided summarisation track at the 2010 Text Analysis Conference (TAC). The dataset is composed of approximately 44 topics, divided into five categories: Accidents and Natural Disasters, Attacks, Health and Safety, Endangered Resources, Investigations and Trials. Each topic has a topic ID, category, title, and 20 relevant documents which have been divided into 2 sets: Document Set A and Document Set B. Each document set has 10 documents, and all the documents in Set A chronologically precede the documents in Set B. Originally in the TAC task, given a topic 2 summaries should be generated one for Document Set A and one for Set B. The summary for set B should avoid the repetition of information already found in the topic's set A. For our experiments we use only the documents in Set A for each topic. The summarisation data was organised such that each topic had 4 human summaries manually written by humans against which the automatic summaries were evaluated.

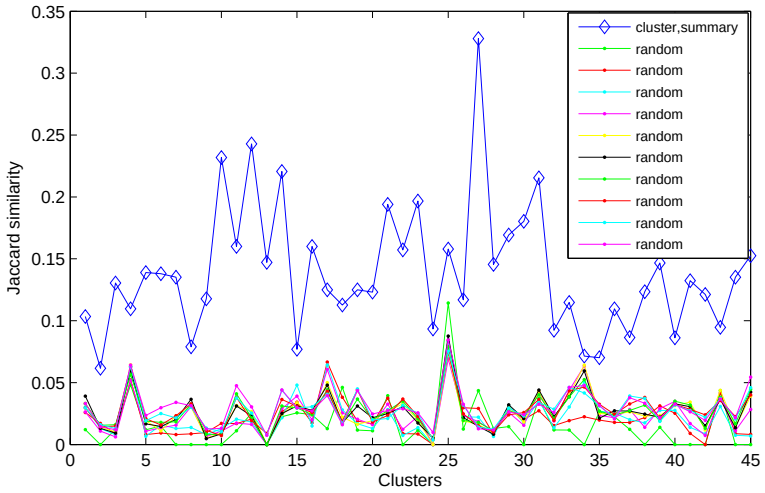
5.1.2 Preliminary Work

Before extracting the key entities and actions from the summarization dataset we wanted to benchmark our approach using the human summaries (which are manually written) that are available for each topic in the dataset. We prove that the key entities and actions extracted using our pipeline from both the documents(dataset) and the corresponding human summaries for each topic are significantly correlated and therefore using the key entities and actions as features in generating automatic summaries would be advantageous. The Jaccard similarity coefficient (Manning and Schütze, 1999) $J(X, Y)$ is used to measure this similarity that is defined as follows:

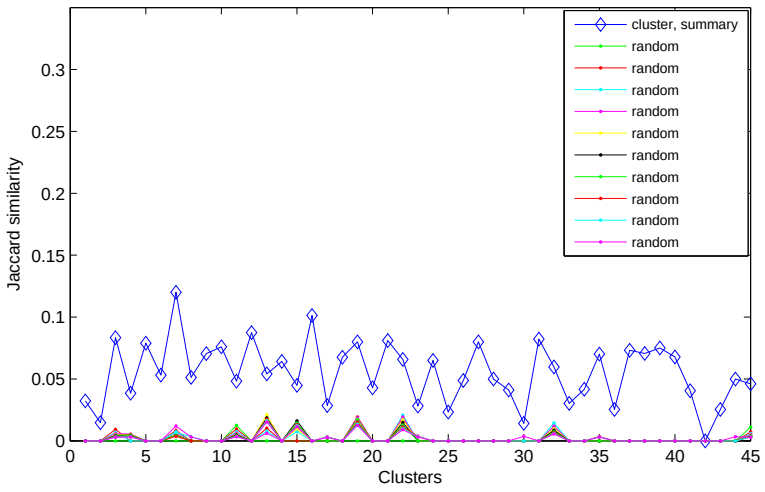
$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{5.1}$$

where X is the set of key entities extracted from the documents in a topic and Y is the set of key entities extracted from the human summaries related to that topic. Key entities are defined as the most frequent entities identified in text or the highest weighted entities according to Equation 2.1 discussed in Chapter 2. We explore the effect of these two settings, considering the top 50 highly frequent actors and highly weighted actors identified, in measuring the Jaccard similarity of entities between the documents and human summaries for each of the 44 topics. As a measure to reject the null hypothesis we measure similarity between key entities from human summaries and key entities extracted from the documents selected from a random topic out of the 44.

Figure 5.1a and 5.1b show the Jaccard similarity coefficient between summary/topic-entities and in 10 random settings (summary/random topic-entities). Figure 5.2a and 5.2b show the same for the top 50 most frequent actions and weighted actions identified.

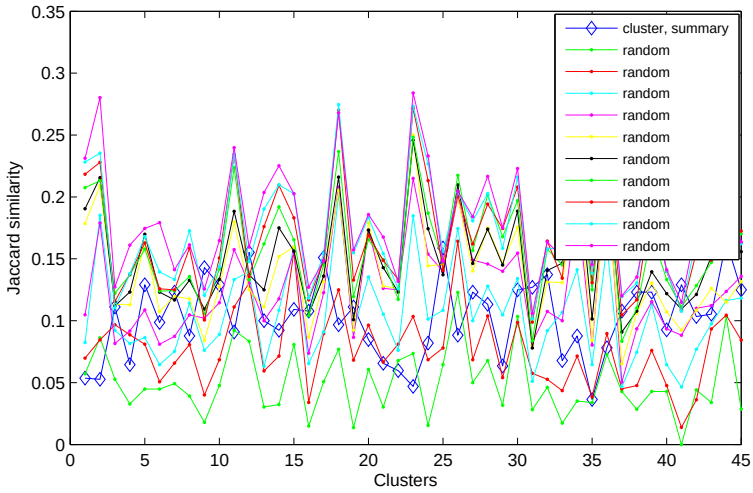


(a) Most frequent entities

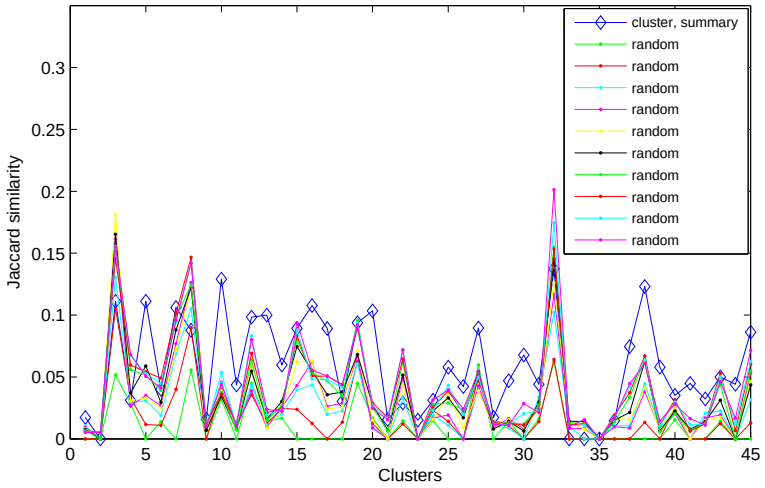


(b) Weighted entities

Figure 5.1: Jaccard similarity plot between cluster-summary and random for the top 50 most frequent entities and weighted entities.



(a) Most frequent actions



(b) Weighted actions

Figure 5.2: Jaccard similarity plot between cluster-summary and random for the top 50 most frequent actions and weighted actions.

However it is evident from figures 5.1a and 5.1b that both the highly frequent entities and the highest weighted entities correlate very well with the summary entities. In the case of actions the weighted actions show better correlation than the most frequent actions. Based on this we assume that key sentences that form a good summary should contain the most frequent/highest weighted entities and the

highly weighted actions. However we found later that using the highest weighted entities produced better summaries than using the most frequent entities.

5.1.3 Methodology

5.1.3.1 Pre-processing

Each document in the document set contains tags, such as <DOC>, <DOCTYPE>, </DOCTYPE>, <HEADER>, </HEADER>. We extract the text between <TEXT>, </TEXT> tags. Once the text is extracted we split them into sentences and store them. We ignored sentences that contained dialogues enclosed within double quotes and also sentences that started with stop words such as 'One', 'Two', 'But' and 'That'. For each topic we extract they key entities, key actions and SVO triplets using our pipeline. We also incorporated the Stanford POS tagger and NP Chunker into our pipeline to extract noun phrases in the text. With this we were able to collect the most frequent noun phrases for each topic as well.

5.1.3.2 Extraction of Summary Sentences

The next step in our method was to identify a subset from the whole list of sentences that are eligible to go into the summary. In order to do this we first extracted from the list of triplets we had in total for each topic, only the triplets that contained the 50 highest weighted key entities and key actions. We call them the key triplets.

A search was performed on the sentences in document set A relevant to a topic based on the key triplets. Sentences that contained the subject, verb and object elements in the triplet in the right chronological order were extracted. For each triplet we were able to see many semantically equivalent sentences. The shortest sentence was selected in that case. Finally we had a list of sentences that corresponded to all the key triplets. From these we extracted the most relevant summary sentences using a sentence scoring method.

5.1.3.3 Sentence Scoring

The candidate sentences were assigned a score based on three features, number of key entities (na), number of key actions (nb) and the most frequent noun phrases (nc) that were found in the sentence. The final score S for a sentence was computed using the scoring function,

$$S = \alpha(S1) + \beta(S2) + \gamma(S3) \tag{5.2}$$

where $S1{=}na/S_{len}$, $S2{=}nb/S_{len}$ and $S3{=}nc/S_{len}$. S_{len} is the length of the sentence. This leads to the following optimisation problem,

$$argmax_{\alpha\in\{0,1\},\beta\in\{0,1\}}S \tag{5.3}$$

for which S reaches its largest value R for parameters α, β and γ. R is the ROUGE Score obtained for the automatically generated summary. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. Formally, ROUGE-N is a n-gram recall between a candidate summary and a set of reference summaries. It is computed as follows,

$$ROUGE-N = \frac{\sum_{S\in ReferenceSummaries}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in ReferenceSummaries}\sum_{gram_n\in S} Count(gram_n)} \tag{5.4}$$

where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries (Lin, 2004). In the TAC summarization task ROUGE-2 , i.e. N=2 (bigrams) and ROUGE-SU4 are being used to evaluate the summaries. ROUGE-SU is based on skip-bigrams which is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics measure the overlap of skip-bigrams between a candidate summary and a set of reference summaries. ROUGE-SU4 looks for matching bigrams with skip distance up to 4 words, stemmed (Lin, 2004). We use the ROUGE score to evaluate our summarization system.

There were altogether six parameters that needed to be identified which maximized the ROUGE scores. They are α, β, γ and $t1$, $t2$ and $t3$ which are the

thresholds for na, nb and nc.

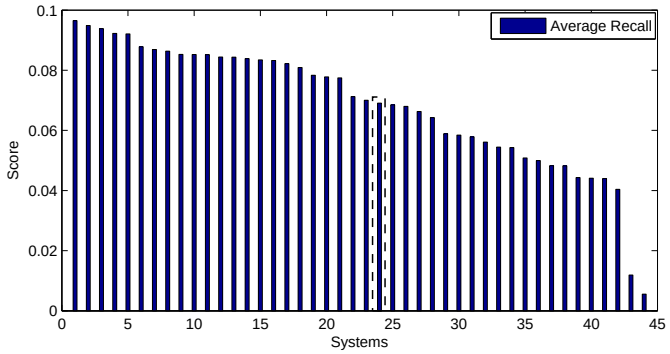
The parameters α, β were assigned values between 0 and 1 with 0.1 interval for calculating the ROUGE scores where $\gamma = 1 - \alpha - \beta$. Thresholds $t1$, $t2$ and $t3$ were assigned values from 10 to 50. The six parameters were selected after repeatedly computing ROUGE (ROUGE-2 and ROUGE SU4) scores for the values assigned to them and selecting the ones that produced higher scores. The optimal parameters that maximised the score were $\alpha = 0.1$, $\beta = 0.4$, $\gamma = 0.5$, $t1 = 25$, $t2 = 25$ and $t3 = 25$. We used the same thresholds and parameters for all our summarization experiments.

Sentences were ranked according to their scores S and the least scoring sentence was removed iteratively until the remaining sentences made up a 100 word summary which is the requirement of the TAC guidelines. If there is more than one sentence with the same least score, we always retain the longest sentence out of them. We add it to the final summary if the summary word count becomes much less than 100.

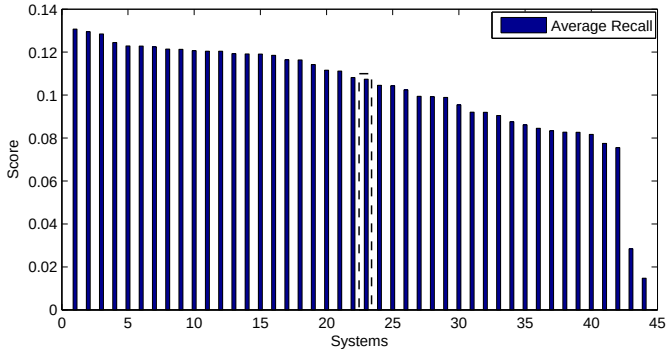
5.1.3.4 Evaluation

The summarization system developed by us was evaluated by comparing our ROUGE scores with that of those who participated in the 2010 TAC summarization task. The scores of the participants was readily available in the TAC evaluation results for comparison. Figure 5.3 shows the ROUGE-2 and ROUGE-SU4 average recall scores of all systems participated in the TAC 2010. Our system ranks 24 in ROUGE-2 recall and 23 in ROUGE-SU4 recall out of the 44 systems.

Therefore the key entities, actions and triplets identified in text proves to be a useful method for the summarization task. Future work could be conducted in order to generate much more cleaner summaries using this information with other more sophisticated techniques. In the guided summarisation task the test data is divided in to five categories like discussed before and participants (and human summarizers) are also given a list of important aspects for each category which the summary should cover ideally. For example for the topic “Accidents and Natural Disasters" the following aspects are provided.



(a) ROUGE-2 recall



(b) ROUGE-SU4 recall

Figure 5.3: ROUGE-2 and ROUGE-SU4 average recall results for 44 systems in TAC 2010.

- WHAT: what happened
- WHEN: date, time, other temporal placement markers
- WHERE: physical location
- WHY: reasons for accident/disaster
- WHO AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the accident/disaster
- DAMAGES: damages caused by the accident/disaster
- COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the accident/disaster

We have not used these aspects in our system. It would be useful to make use of them by detecting them and forcing sentences containing the aspects to

be in the summary. We could also make use of other features from text such as position of key entities/actions, triplets in text.

5.1.4 Summary

This section discussed extensively about extractive summarization which is used in automatic summarization in recent times. Automatic multi-document summarization was shown to be one of the key applications of triplets identified from text data. We discussed a method to do automatic summaries using the key entities, actions and triplets found by our system pipeline incorporating a sentence scoring method based on textual features and using it to identify the summary sentences. This method proves to be useful by ranking 24 and 23 in ROUGE-2 and ROUGE-SU4 recall scores when compared across the scores of other 43 participants in the TAC 2010 guided summarization task.

5.2 Further Applications

5.2.1 Question and Answering Systems

Structured information is becoming increasingly important for many applications and one of them is providing answers to questions in natural language. Among various other approaches, triplets and semantic networks generated from triplets have been used for question and answering. Dali et al. (2009) presents a system that provides answers to questions browsing through the document that supports the answer. The questions follow a predetermined template, whereas the answers are yielded based on the previously extracted information, in the form of subject-verb-object triplets. The questions' parse trees undergo linguistic analysis to determine the type of the question and to translate it to a query for the triplet search engine. Once the answers are found in the form of triplets, an explanation is provided for each of them, in the form of triplets, sentences and documents it was derived from.