Generating Headline Summary from a Document Set

Kamal Sarkar and Sivaji Bandyopadhyay

Computer Science & Engineering Department, Jadavpur University,
Kolkata – 700 032, India
jukamal2001@yahoo.com, sivaji_ju@vsnl.com

Abstract. This paper discusses an approach to generate headline summary from a set of documents. Headline summary is basically a very short summary in the form of headline. As the amount of on-line information increases, systems that can automatically summarize multiple documents are becoming increasingly desirable. In this situation, headline summary is useful for users who only need information on the main topics in a set of documents. Headline summary from multiple documents will be very useful in the text mining applications for the generation of meaningful label (a compact identifier that allows a person to quickly see what the topic is about) for a cluster of documents.

1 Introduction

In this paper we present a system that will cluster the text documents collected from multiple online sources and generate a headline summary in one or two sentences for each cluster by identifying named entities from each document in the set to make a global list of named entities and forming a headline summary for the set.

All the previous work on headline generation [1, 2] was done on single documents but the focus in the present work is on headline summary generation from a set of documents. Moreover, instead of using only statistical approaches, we have used named entity cues and summary generation techniques for our work, since it is very difficult to have a training corpus of document set—headline pairs. In the next section we present the proposed approach. The system is evaluated in Section 3.

2 Proposed Approach

News collected from multiple sources should be clustered. Clustering technique adopted is similar to the method used by Chen and Lin [3].

The input to our system is a cluster of related documents. Based on the observations of human-produced headline summary, we have developed the following algorithm.

2.1 Algorithm

1. Prepare the local named entity list for each document in the cluster.
2. Prepare the global named entity list for the cluster.
   - Resolve the cross-document co-reference by the approach used in [4]. It uses a global translation table, changing all occurrences of each co-referred named entity to the longest version.

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 649 – 652, 2005.

- Rank the named entities according to their frequency across the local named entity lists.
- Take top *n* named entities in the global list. Value of n should be chosen in such a way that total number of words in the selected named entities should not exceed 10 words, because our objective is to generate the very short summary.

3. Pair the named entities in the global list, which co-occur in the same sentence and one of them occurs in the subject position of the sentence.
4. Identify meaningful sentence segments from the documents in the cluster. If no pairing is possible in Step 3, the sentences containing the most frequently named entity are selected. Otherwise, the following rules are used in the specified order.

   Rule1: *For each named entity pair (A,B) identified in Step 3, select the sentence segment from A to B including both A and B.*
   Rule2: *If (A,B) and (B,C) are two pairs occurring in the same sentence, and A>B>C (A>B means A occurs before B), the sentence segment from A to C is selected.*
   Rule3: *For the rest of the named entities that occur in the subject position and are not a member of any pair, pair each of the named entities with the main verb/verb group of the sentence concerned.*
   Rule4: *The named entities that do not participate in a pair and do not occur in the subject position of any sentence in the document set should simply be ignored.*

5. Headline summary generation:

   The meaningful sentence segments selected for each of the pairs would be clustered. The sentence segments selected using rule 3 in the Step 4 would be clustered separately for each such named entity. If no pairing is at all possible, only one cluster is formed.

   Each cluster will contribute one representative to the final headline summary. If a cluster contains more than one sentence segment, the most summarized one, i.e., the segment containing least number of words will be the representative from that cluster.

   Finally, all the representative sentence segments from the clusters are arranged in a particular order called majority ordering [5], which relies on the original order of sentences in the input documents where from the sentence segments have been selected.

2.2 Example

The following 5 documents have been collected from the different newspapers on US Space Shuttle Columbia crash. It has been illustrated with this example how our algorithm works on this cluster to generate a headline summary in a few sentences.

Doc-1: <Title> Columbia crashes on return <Title>
<Text Start> *Kalpana Chawla, who traveled more than any other Indian in history, met with a tragic, fiery end to her life when American space shuttle, Columbia, in which she was an astronaut, broke up and crashed over Texas* only minutes before it was to land in Florida. Israeli

first astronaut, Ilan Ramon, an air force pilot, perished in the tragedy along with Chawla and five other crew members of the ill-fated above flight.<Text End>

Doc-2: <Title> Space shuttle explodes <Title>
<Text Start> US Space shuttle Columbia with Indian-orgin astronaut *Kalpana Chawla on board burst into flames over Texas,* killing all seven astronauts, minutes before it was to land in Florida on Saturday. In North Texas, several residents reported hearing a big bang, the same time when all radio and data communication with the shuttle and its crew was lost.<Text End>

Doc-3: <Title> Columbia explodes <Title>
<Text Start>*Space shuttle Columbia broke apart in flames as it streaked over Texas* towards its scheduled landing, killing, this time too, all seven astronauts on board, six Americans including India-born Kalpana Chawla, and Ilan Ramon, the first Israeli astronaut to go into space. Nasa did not immediately declare the crew dead, but the *US* flag next to its countdown clock was lowered to half-staff. Later in the day, President Bush said US space exploration would continue despite the loss.<Text End>

Doc-4: <Title> Chawla, rest of crew killed on board space shuttle <Title>
<Text Start>US space shuttle Columbia, carrying seven astronauts including Indian-American Kalpana Chawla, disintegrated shortly before landing at *Cape Canaveral* on Saturday morning in what appeared to be a ghastly replay of the Challenger disaster 27 years ago. NASA sources have confirmed that death of all seven on board.<Text End>

Doc-5: <Title> Columbia burns up over Texas <Title>
<Text Start>*The space shuttle Columbia disintegrated over the state of Texas* today, minutes before its scheduled landing in Florida, killing all seven astronauts on board. Six of them were *Americans* including the Indian-American, Kalpana Chawla and one Israeli air force officer. Calling it as indeed a tragic day for the NASA family, the top administration of the agency, Seon Keefe told a press briefing that the terrible tragedy was not caused from the ground.<Text End>

In the documents, the local named entities have been shown by the underlined words. In the example, the cross document co-references like *US Space Shuttle, US Space Shuttle Columbia, Space Shuttle Columbia, American Space Shuttle, The Space Shuttle Columbia can be resolved to the longest version, US Space Shuttle Columbia*.

Global named entity list: *Kalpana Chawla, US Space Shuttle Columbia, Texas*.
Named Entity Pairs: (*Kalpana Chawla, Texas*), (*US Space Shuttle Columbia, Texas*).
Selected sentence segments: The italic sentence segments in the above documents are the meaningful sentence segments identified by our approach.
Resulting headline summary: Cluster#1 for the pair (*Kalpana Chawla, Texas*) and Cluster#2 for the pair (*US Space Shuttle Columbia, Texas*) will contribute the following sentences to the final summary: *The space shuttle Columbia disintegrated over the state of Texas. Kalpana Chawla on board burst into flames over Texas.*

3 Evaluation and Results

We have extracted the set of distinct words from the headlines of the documents in the document set and this set of words has been considered as a gold standard. While

comparing the output summary and the gold standard, we have considered whether they are string identical or synonymous. The precision and recall have been computed as follows.

If the reference headline summary (words in the gold standard) is of length n words, the generated headline summary is of length k words and p of n words are in the generated headline summary, Precision = p/k and recall = p/n.

For the small document set (< 10 documents) we used the above evaluation method. However, if the size of the document set is very large (say, 100 documents), the reference headline summary is likely to increase in size. So, we have restricted the size of the reference headline summary to the size of generated summary by selecting words from the list of compiled headline words by their frequencies across the headlines of the documents.

For 5 sets of documents collected from the newspapers, the average precision and the average recall which have been achieved by our system are 0.51 and 0.645, respectively.

4 Conclusion and Future Work

We have presented an algorithm for generating a headline summary from a set of related documents. The final summary can be evaluated by an extrinsic evaluation method that determines how well the machine-generated summary can classify the document sets in the test corpus. This evaluation method will be investigated in future.

References

1. Dorr, B., Zajic, D., Schwartz, R.: Hedge trimmer: a parse-and-trim approach to headline generation. In *Proceedings of Workshop on Automatic Summarization* (2003).
2. Zhou, L., Hovy, E.: Headline Summarization at ISI. In proceedings of Workshop on text summarization. Edmonton, Canada (2003).
3. Chen, H., Lin, C.: A multilingual news summarizer. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 159–165 (2000).
4. Clifton, C., Cooley, R.: TopCat: Data Mining for Topic Identification in a Text Corpus, In IEEE transactions on knowledge and data engineering. Vol.16, No.8, pages 949–964 (2004)
5. Barzilay, D., Elhadad, N., McKeown, R. K.: Sentence Ordering in Multi-document Summarization. *In Proceeding of Human Language Technology Conference (HLT), San Diego (2001).*