E-037

# Multi-Document Summarization using BE-breaking

Md. Waliur Rahman Miah
wali@nlp.is.ritsumei.ac.jp
Graduate School of Science and Engineering
Ritsumeikan University

Junichi Fukumoto
fukumoto@media.ritsumei.ac.jp
Department of Media Technology
Ritsumeikan University

## Abstract

We have used Basic Element (BE) Breaking method to generate a better summary in multidocument environment. We count the content level information in sentences among multiple documents to find out the similar parts. Truncating the redundant similar parts and merging the others would improve the quality of automated multidocument summary.

## 1. Introduction

In multiple document summarizations a system ranks the sentences among documents and then collects important ones in a summary. In such systems it happens that sentences containing similar information but syntactically different are collected due to there high individual scores. Also the sentences containing partial important information may be discarded. In our present research we focused on the content level information of each individual sentence. We find out content similarity to avoid accumulation of semantically similar sentences. We also looked for partial information contents in the sentences and merge them so that the final summary becomes rich with information.

To find the content level information similarity in a sentence, a mechanism called Basic Elements (BE) was proposed by Hovy, et.al.[1]. BEs consists of two elements (head and modifier) and relation between those two elements (head-modifier-relation). A BE is generated through syntactic analyzer as a minimal semantic unit in a sentence. Evaluation of document summaries is done by comparison between system summary and human summary. Fukumoto et.al.[2][3] utilized this mechanism for automatic evaluation of Question Answering system.

We also use BE method for detecting similar part of sentences to generate a summary because BE works as minimal-length fragments of 'sensible meaning' and may give content level information of sentence parts. Finding the content information, we can discard the redundant parts and merge the important ones to make a compressed summary with better information contents.

## 2. System Design

In this research, we have utilized BE matching techniques to generate a better summary. We calculated the similarities among the sentences between two documents, picked up one between the two similar sentences and stored the additional elements of the sentence which has been discarded.

Basic algorithm is as follows:

1. Make preliminary individual summaries.
2. Calculate similarities among sentences between two documents using BE method.
3. Identify core elements and additional elements
4. Marge them.

## 3. Implementation Methodology:

At first we make the preliminary individual summaries through tf-idf based summarizer[4]. We tried one of these kinds of summarizer named MEAD from http://tangra.si.umich.edu/clair/md/demo.cgi

In the tf-idf method, longer sentences get greater scores, hence are extracted easily, but the important smaller sentences are discarded due to smaller scores. To solve this problem we corrected the preliminary individual summaries for the analysis of BE matching by hand. We have used BE breaker to get the BEs of sentences. BE breaker Package is available without restriction at http://haydn.isi.edu/BE/.

## 4. Experiment and Result

### 4.1 Experiment

To test and analyze our system we have used three sets of files each set consisting two files of similar topics. Due to the space constraint of this paper we illustrate two sentences from two different files of a set. The files have been used to make a summary.

First we have broken every sentence into their BEs by using the BE Breaker.

```
Filename: yunus1
Sentence No.: 2
He was born on June 28, 1940 in
Chittagong in Bangladesh.        ...[S1]
```

BEs:

```
 1: be <- born (pred)
 2: born <- june 28 , 1940 (on)
 3: be <- chittagong (in)
 4: chittagong <- bangladesh (in)
```

```
Filename: yunus2
Sentence No.: 2
The third oldest of nine children,
Yunus was born on June 28, 1940 to a
Muslim family in the village of Bathua,
by the Boxirhat Road at Hathazari in
Chittagong in Bangladesh.            ...[S2]
```

BEs:

```
 1: third <- old (pnmod)
 2: children <- nine (nn)
 3: old <- children (of)
 4: be <- born (pred)
 5: born <- june 28 , 1940 (on)
 6: family <- muslim (nn)
 7: be <- family (to)
 8: family <- village (in)
 9: village <- bathua (of)
10: be <- boxirhat road (by)
11: boxirhat road <- hathazari (at)
12: hathazari <- chittagong (in)
13: chittagong <- bangladesh (in)
```

After the BE breaking we find out the similar BEs in each and every pair of sentences between the two files. In the above example sentences the similar part and similar BEs are underlined.

Overall similarity between two sentences is measured by BE-similarity as follows:

$$\text{BE-Sim (S1,S2)} = \frac{\sum MatchedBEs}{\sqrt{\sum BEs(S1)}.\sqrt{\sum BEs(S2)}}$$

For the above particular example:
Overall BE-similarity between the two sentences S1 and S2 is:

$$\text{BE-Sim(S1,S2)} = \frac{3}{\sqrt{4}\cdot\sqrt{13}} = \frac{3}{7.21} = 0.416$$

We calculate the shared ratio (SR) of BE-similarity of one sentence in between the two as follows:

Shared ratio of BE-similarity of a sentence:

$$\text{SR}=\frac{No.ofSimilarBEs}{No.ofTotalBEs}$$

In the above two sentences S1 and S2:
Shared ratio of BE-similarity of S1 is;

$$\text{SR(S1,S2)} = \frac{3}{4} = 0.75$$

Shared ratio of BE-similarity of S2 is;

$$\text{SR(S2,S1)} = \frac{3}{13} = 0.23$$

### 4.2 Result

From the experiment in the Shared ratio of BE-similarity we see that less part (0.23) of 2nd sentence is same as most part (0.75) of 1st sentence i.e. the 2nd sentence is bigger and contains most part of the 1st sentence. In this situation if we pickup the 1st sentence and discard the 2nd then we will have to deal with more unmatched parts from the 2nd sentence. On the other hand if we take the 2nd sentence then most part of 1st sentence is been taken within it and less unmatched parts of 1st sentence will have to be merged. Therefore we take the 2nd sentence and extract the unmatched BEs from the 1st sentence. In this particular case, the partial unmatched BE is the 3rd BE of 1st sentence.

In this way by comparing each and every sentence between two files we find out the similarity, then select the bigger sentence and extract the dissimilar part of unselected one. If there is no similarity between the two sentences then both of them are selected.

### 5. Discussion

We succeed to make a summary from multiple documents of same topic by comparing the sentences between the documents and extracting the important ones. A central motivation for BEs is that each piece of information can be counted. Then in the summary, important information can be included and redundant information can be excluded by finding out similarities between BEs. Though we have succeeded to find similarities between many sentences through their BEs but there are still some sentences and part of sentences which are semantically same but could not recognized by the BE system. Observation shows that if some words are replaced by there synonyms then they produce similar BEs. A smart BE matcher that can use dictionaries and thesaurus for more efficient matching could be a further research theme.

We also have collected the partial information from the un- extracted sentences in the form of BEs. This partial information can be merged with the extracted sentences to make the summary a concise one and rich with information. To device the paraphrase rules and design the algorithm for this automatic merging technique is a good theme for future research.

## 6. Conclusion

In this paper we applied BE method for making multi-document summary. We have conducted experiments that compare two sentences between documents. According to BE-based similarity, we could choose important contents among documents and choose one of additional elements. Finally we showed that merging them would make a good informative summary. In the experiment, using news paper articles and biographic articles, we proved that BE method is applicable in summary generation. However there were some cases where BEs did not match though the sentences were describing semantically similar thing but in different lexical or syntax structure. This is because BEs are extracted from the parse tree of sentences by syntactic analysis. To improve the performance of this method it is necessary to develop paraphrase rules for the BE list in order to work in the lexical and syntax level.

## References

[1] E. Hovy, C.Y. Lin, L. Zhou, and J. Fukumoto. 2006. ***Automated Summarization Evaluation with Basic Elements.*** In Proc. of the 5th LREC, Genoa, Italy.

[2] J. Fukumoto, 2007. ***Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method.*** In Proc. of the 6th NTCIR Tokyo, Japan.

[3] A. Yamamoto and J. Fukumoto 2008. ***Automatic Evaluation of Question Answering System based on BE Method.*** In Proc. of the 23rd International Technical Conference on Circuits/Systems, Computers and Communications. Yamaguchi, Japan

[4] E. Hovy and C.Y. Lin. 1999. ***Automated Text Summarization in SUMMARIST.*** In Advances in Automatic Text Summarization, I. Mani and M. Maybury (editors), 1999