Summarizing Answers for Complicated Questions

Liang Zhou, Chin-Yew Lin and Eduard Hovy
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
USA
{liangz, cyl, hovy}@isi.edu

Abstract

This paper describes the multi-document summarization system designed by the Webclopedia team from ISI for DUC 2005. In contrast to the past DUCs and previous designs, this version of our summarizer consists of a query-interpretation component that directly analyzes the given user profile and topic narrative for each document cluster before creating a corresponding summary. This system ranks 4th on ROUGE-1, 7th on ROUGE-2 and ROUGE-SU4. Evaluations conducted by Basis Elements show this system ranks 6th among 32 automatic systems.

1 Introduction

This paper describes a query-based multi-document summarizer based on basic elements (BE) (Hovy et al. 2005), a head-modifier-relation triple representation of document content developed at ISI. BEs are intended to represent the high-informative unigrams, bigrams, and longer units of a text, which can be built up compositionally. An important aspect is that they can be produced automatically. However, BEs can also be used to interpret topic-based queries, as a counting unit for frequency-based topic identification. The idea is to assign scores to BEs according to some algorithms, assign scores to sentences based on the scores of the BEs contained in the sentences, and then apply standard filtering and redundancy removal techniques before generating summaries. Our experimental results show that this approach was very effective in DUC 2005. Figure 1 illustrates the overall system design.

In the following sections, we give a short overview of Basic Elements in the next section. Section 3 describes the BE-based multi-document summarizer. Section 4 shows the performance in DUC and we conclude and discuss future directions in Section 5.

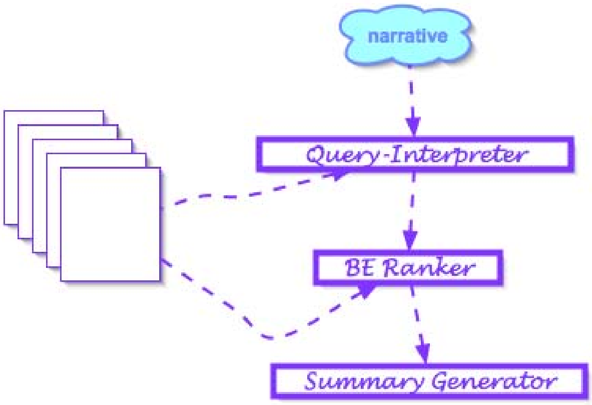


Figure 1. Overall system design.

2 Basic Elements (BE)

At the most basic level, Basic Elements are defined as follows:

- the head of a major syntactic constituent (noun, verb, adjective or adverbial phrases), expressed as a single item, or
- a relation between a head-BE and a single dependent, expressed as a triple (head | modifier | relation).

BEs can be created automatically in several ways. Most of them involve a syntactic parser to produce a parse tree and a set of 'cutting rules' to extract just the valid BEs from the tree.

With BE represented as a head-modifier-relation triple, one can quite easily decide whether any two units match (express the same meaning) or not–considerably more easily than with longer units, of the kind that have been suggested for summarization evaluation by other researchers (Van Halteren and Teufel, 2003; Nenkova and Passonneau, 2004). For instance, "United Nations", "UN", and "UNO" can be matched at this level (but require work to isolate within a longer unit or a sentence), allowing any larger unit encompassing this to accept any of the three variants.

Example BEs for "two Libyans were indicted for the Lockerbie bombing in 1991" are as follows, written as (head | modifier | relation):

<table>
<tr><td>libyans|two|nn</td><td>(BE-F)</td></tr>
<tr><td>indicted|libyans|obj</td><td>(BE-F)</td></tr>
<tr><td>bombing|lockerbie|nn</td><td>(BE-F)</td></tr>
<tr><td>indicted|bombing|for</td><td>(BE-F)</td></tr>
<tr><td>bombing|1991|in</td><td>(BE-F)</td></tr>
</table>

The BEs shown above (BE-Fs) are generated by BE package 1.0 distributed by ISI[1]. We used the sta ndard BE-F breaker included in the BE package in all our experiments described in this paper.

3 Query-based BE summarizer

We modeled our BE-based multi-document summarizer after the very successful NeATS (Lin and Hovy 2002). It includes the following four major stages.

(1) Query Interpretation

In contrast to past DUCs, the summarization task for this year involves creating a fluent summary for a document set in answering a set of complicated questions placed in a narrative along with a user profile. The narrative consists of several questions related to a central topic that may or may not be discussed thoroughly in the given document set. The user profile indicates the granularity of the summaries to be created. There are two different profiles, "general" or "specific". When given the "specific" profile, one is to create summaries that would include individual instances of the general topic described in the narrative.

A generic summarizer, not tuned for answering queries, extracts the most salient information from a set of documents related to a central topic. When a specific query is given, this assumption is no longer valid. We are not to assume the central topic from the document set is the topic that we are to summarize. The query, in this case the narrative, tells us the desired summarizing topic. To interpret the narrative in ways that will concur with the design of a purely BE-based summarizer, we expand the questions from the narrative and then use the expansions to direct the next step, identifying important BEs.

Questions from a narrative were tagged with part-of-speech tags. Using WordNet, we found synonyms for nouns and verbs that occurred in the narrative. To distinguish which words and their expansions are indeed important and uniquely indicative in the document set, we calculate for each word *w* (including expansion words) its informativeness, which is the ratio of the document set *ITF* of *w* and the world *ITF* of *w* (from Wall Street Journal). If a word's informativeness is greater than a preset threshold, then this word is retained in the expansion list.

(2) Identify Important BEs

BEs were used as counting unit. We replaced unigram, bigram, and trigram counting in NeATS with BE-F counting, i.e. breaking each sentence into BEs instead of unigrams, bigrams, and trigrams. We then computed likelihood ratio (LR) for each BE. The LR score of each BE is an information theoretic measure (Dunning, 1993; Lin and Hovy, 2000) that represents the relative importance in the BE list from the document set that contains all the texts to be summarized. Sorting BEs according to their LR scores produced a BE rank list. We then used the expansion list described in previous section to further filter out this ranked BE list.

(3) Identify Important Sentences

The score of a sentence is the sum of its BE scores computed in (1) divided by the number of BEs in the sentences. We call this normalized sentence BE score. Sorting sentences according to their normalized sentence BE scores produced a ranked list of sentences. By limiting the number of top BEs that contribute to the calculation of sentence scores, we can remove BEs with little importance and sentences with many less important BEs. We call this parameter B. For example, B = 64 means that only the topmost 64 BEs in the rank list created in (1) can contribute to normalized sentence BE score computation.

(4) Generate Summaries

The easiest way to create summaries from (2) is just to output the topmost N sentences until the required summary length limit. However, this simple approach does not consider interactions among summary sentences, such as redundancy and coherence. For example, we should only include one of two very similar sentences with high normalized sentence BE scores in a summary. Goldstein et al. (1999) observed this in what they called maximum marginal relevancy (MMR). This we modeled by BE overlap between an intermediate summary and a to-be-added candidate summary sentence. We call this overlap ratio R, where R is between 0 and 1 inclusively. For example, R = 0.8 means that a candidate summary sentence, *s*, can be added to an intermediate summary, *SI*, if the sentence has a BE overlap ratio less than or equal to 0.8.

Also, given the importance in the news genre of sentence position (Lin and Hovy, 1997), we would like to model the position preference that favors sentences appearing earlier in a document. This is controlled by parameter N. For example, N = 10 means that only the first 10 sentences in a document can be considered as candidate summary sentences.

In favor of leading sentences of the news genre and provide a simple way to improve coherence (lead sentences usually give the setting of news events), we adopted a first-sentence-priority policy, i.e. if a to-be-added candidate summary sentence is not a lead sentence and its lead sentence[2] is yet not included in the immediate summary, then add its lead sentence first when its addition does not violate the overlap ratio

[1] BE website: http://www.isi.edu/~cyl/BE

[2] The lead sentence of a document is the lead sentence of all the sentences in the document.

constraint. This strategy was used with considerable success in NeATS.

Through experimentation using the DUC 2003 task 2 corpus, we found that the BE-based multi-document summarizer with B = 64, R = 0.8, and N = 10 achieved a BE-F score of 0.0532 that was better than the summaries generated by NeATS (at 0.0503) in DUC 2003. We therefore decided to use this set of parameters in DUC 2005.

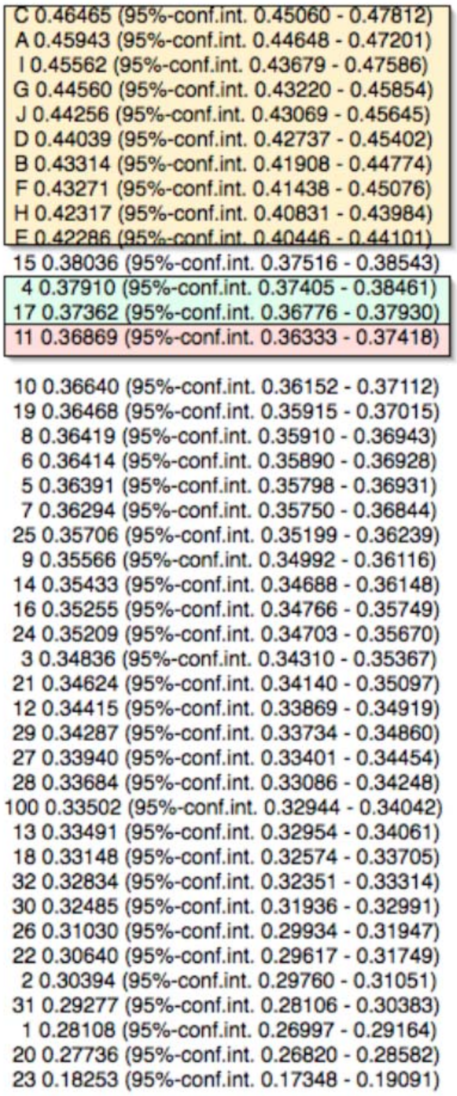
4 Results

Among 32 automated systems, our system performed well on multiple evaluation metrics. System "1" is the baseline which is taking the first 250 words from the most recent article in a particular document set. Our system is identified as "11". Table 1 shows the overall system ranking on ROUGE-1 recall (Lin, 2004). Our system ranks 4th and statistically is not significantly different from systems "4" and "17" ranked 2nd and 3rd respectively. The first highlighted box (yellow) shows how humans perform at a much more superior level. Table 2 shows the scores and the ranking produced by BE with HM (head-modifier) parameterization (Hovy et al., 2005).

Post-DUC Experiments

Without query-interpretation (QI) component, our system outperforms all other systems evaluated by BE on DUC 2002 and 2003 data. After the official DUC 2005 results was released, we wanted to see whether this QI component is effective and by how much. Is there any difference between generic summaries and query-based summaries? We duplicated the settings used in the automatic evaluation metrics ROUGE and BE. In Table 1and 2, system "100" is our system without the QI component. As we can see, generic summaries produced on the document sets are not suitable answers for the questions raised in the narratives. Using BEs to interpret the narratives and to subsequently select top BEs produces adequate summaries for query-based summarization tasks.

One of our pre-DUC suspicion is that sentence compression, especially an MMR-based compression strategy (Hovy et al., 2005), would not contribute significantly as it did in MSE2005. The primary reasoning for this projected phenomenon is that the given document set is diverse in the topic being asked. More specifically, when we try to address the issues raised in the topic narrative, where questions are asked, we find many and/or large segments of the documents from the set are not relevant. This led us to believing retaining the redundancy in extracted sentences would actually make the summaries better. Since we need to mine among diverse irrelevant information, the redundant segments of sentences add relevance to those sentences and therefore would contribute to the overall summary. Our hypothesis was confirmed by running post-DUC experiments. MMR-based sentence compression results in lower ROUGE and BE scores.



C	0.46465	(95%-conf.int. 0.45060 - 0.47812)
A	0.45943	(95%-conf.int. 0.44648 - 0.47201)
I	0.45562	(95%-conf.int. 0.43679 - 0.47586)
G	0.44560	(95%-conf.int. 0.43220 - 0.45854)
J	0.44256	(95%-conf.int. 0.43069 - 0.45645)
D	0.44039	(95%-conf.int. 0.42737 - 0.45402)
B	0.43314	(95%-conf.int. 0.41908 - 0.44774)
F	0.43271	(95%-conf.int. 0.41438 - 0.45076)
H	0.42317	(95%-conf.int. 0.40831 - 0.43984)
E	0.42286	(95%-conf.int. 0.40446 - 0.44101)
15	0.38036	(95%-conf.int. 0.37516 - 0.38543)
4	0.37910	(95%-conf.int. 0.37405 - 0.38461)
17	0.37362	(95%-conf.int. 0.36776 - 0.37930)
11	0.36869	(95%-conf.int. 0.36333 - 0.37418)
10	0.36640	(95%-conf.int. 0.36152 - 0.37112)
19	0.36468	(95%-conf.int. 0.35915 - 0.37015)
8	0.36419	(95%-conf.int. 0.35910 - 0.36943)
6	0.36414	(95%-conf.int. 0.35890 - 0.36928)
5	0.36391	(95%-conf.int. 0.35798 - 0.36931)
7	0.36294	(95%-conf.int. 0.35750 - 0.36844)
25	0.35706	(95%-conf.int. 0.35199 - 0.36239)
9	0.35566	(95%-conf.int. 0.34992 - 0.36116)
14	0.35433	(95%-conf.int. 0.34688 - 0.36148)
16	0.35255	(95%-conf.int. 0.34766 - 0.35749)
24	0.35209	(95%-conf.int. 0.34703 - 0.35670)
3	0.34836	(95%-conf.int. 0.34310 - 0.35367)
21	0.34624	(95%-conf.int. 0.34140 - 0.35097)
12	0.34415	(95%-conf.int. 0.33869 - 0.34919)
29	0.34287	(95%-conf.int. 0.33734 - 0.34860)
27	0.33940	(95%-conf.int. 0.33401 - 0.34454)
28	0.33684	(95%-conf.int. 0.33086 - 0.34248)
100	0.33502	(95%-conf.int. 0.32944 - 0.34042)
13	0.33491	(95%-conf.int. 0.32954 - 0.34061)
18	0.33148	(95%-conf.int. 0.32574 - 0.33705)
32	0.32834	(95%-conf.int. 0.32351 - 0.33314)
30	0.32485	(95%-conf.int. 0.31936 - 0.32991)
26	0.31030	(95%-conf.int. 0.29934 - 0.31947)
22	0.30640	(95%-conf.int. 0.29617 - 0.31749)
2	0.30394	(95%-conf.int. 0.29760 - 0.31051)
31	0.29277	(95%-conf.int. 0.28106 - 0.30383)
1	0.28108	(95%-conf.int. 0.26997 - 0.29164)
20	0.27736	(95%-conf.int. 0.26820 - 0.28582)
23	0.18253	(95%-conf.int. 0.17348 - 0.19091)

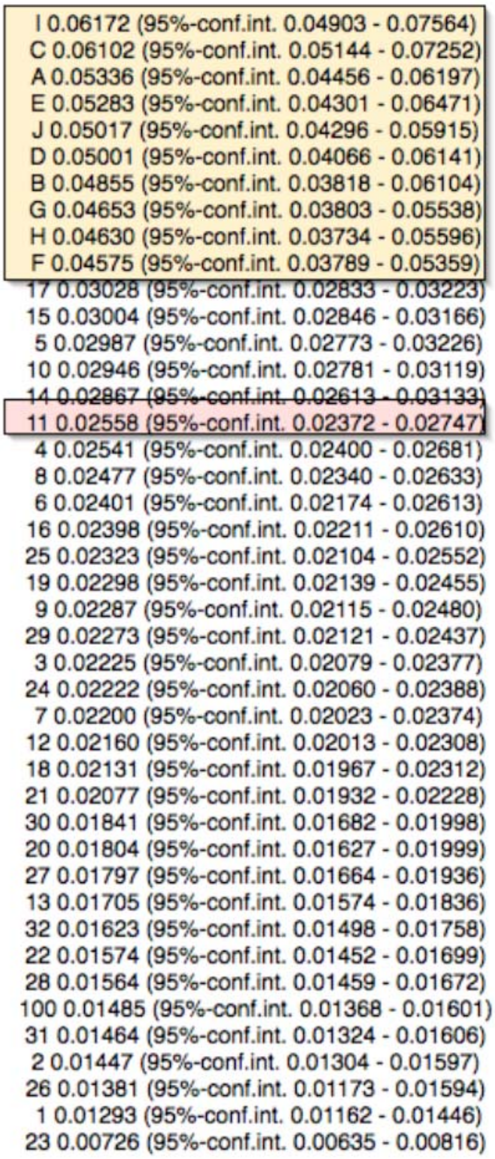
Table 1. ROUGE-1 scores.

5 Conclusion and Future Work

In this paper, we have described a multi-document summarization system that was designed based on the fundamentals of Basic Elements. Through experimentation, we see that a query-interpretation component is critical in addressing summarization need for topic-based tasks. Future work is planned in designing mechanisms that would benefit both query understanding and its linkage to summary creation/extraction.

References

Barzilay, R., McKeown, K., and Elhadad, M. 1999. Information Fusion in the Context of Multi-Document Summarization. *Proceedings of the 37th*



```
  I 0.06172 (95%-conf.int. 0.04903 - 0.07564)
  C 0.06102 (95%-conf.int. 0.05144 - 0.07252)
  A 0.05336 (95%-conf.int. 0.04456 - 0.06197)
  E 0.05283 (95%-conf.int. 0.04301 - 0.06471)
  J 0.05017 (95%-conf.int. 0.04296 - 0.05915)
  D 0.05001 (95%-conf.int. 0.04066 - 0.06141)
  B 0.04855 (95%-conf.int. 0.03818 - 0.06104)
  G 0.04653 (95%-conf.int. 0.03803 - 0.05538)
  H 0.04630 (95%-conf.int. 0.03734 - 0.05596)
  F 0.04575 (95%-conf.int. 0.03789 - 0.05359)
 17 0.03028 (95%-conf.int. 0.02833 - 0.03223)
 15 0.03004 (95%-conf.int. 0.02846 - 0.03166)
  5 0.02987 (95%-conf.int. 0.02773 - 0.03226)
 10 0.02946 (95%-conf.int. 0.02781 - 0.03119)
 14 0.02867 (95%-conf.int. 0.02613 - 0.03133)
 11 0.02558 (95%-conf.int. 0.02372 - 0.02747)
  4 0.02541 (95%-conf.int. 0.02400 - 0.02681)
  8 0.02477 (95%-conf.int. 0.02340 - 0.02633)
  6 0.02401 (95%-conf.int. 0.02174 - 0.02613)
 16 0.02398 (95%-conf.int. 0.02211 - 0.02610)
 25 0.02323 (95%-conf.int. 0.02104 - 0.02552)
 19 0.02298 (95%-conf.int. 0.02139 - 0.02455)
  9 0.02287 (95%-conf.int. 0.02115 - 0.02480)
 29 0.02273 (95%-conf.int. 0.02121 - 0.02437)
  3 0.02225 (95%-conf.int. 0.02079 - 0.02377)
 24 0.02222 (95%-conf.int. 0.02060 - 0.02388)
  7 0.02200 (95%-conf.int. 0.02023 - 0.02374)
 12 0.02160 (95%-conf.int. 0.02013 - 0.02308)
 18 0.02131 (95%-conf.int. 0.01967 - 0.02312)
 21 0.02077 (95%-conf.int. 0.01932 - 0.02228)
 30 0.01841 (95%-conf.int. 0.01682 - 0.01998)
 20 0.01804 (95%-conf.int. 0.01627 - 0.01999)
 27 0.01797 (95%-conf.int. 0.01664 - 0.01936)
 13 0.01705 (95%-conf.int. 0.01574 - 0.01836)
 32 0.01623 (95%-conf.int. 0.01498 - 0.01758)
 22 0.01574 (95%-conf.int. 0.01452 - 0.01699)
 28 0.01564 (95%-conf.int. 0.01459 - 0.01672)
100 0.01485 (95%-conf.int. 0.01368 - 0.01601)
 31 0.01464 (95%-conf.int. 0.01324 - 0.01606)
  2 0.01447 (95%-conf.int. 0.01304 - 0.01597)
 26 0.01381 (95%-conf.int. 0.01173 - 0.01594)
  1 0.01293 (95%-conf.int. 0.01162 - 0.01446)
 23 0.00726 (95%-conf.int. 0.00635 - 0.00816)
```

Table 2. BE-HM scores and ranking.

annual meeting of the Association for Computational Linguistics, 1999, Maryland.

Collins, M. 1997. Three generative, lexicalized models for statistical parsing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97),* 16-23.

Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61-74.

Goldstein, J., M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval* (SIGIR-99), Berkeley, CA, 121-128.

Hovy, E. H., J. Fukumoto, C.-Y. Lin, L. Zhou. 2005. Basic Elements. http://www.isi.edu/~cyl/BE

Hovy, E. H., C.Y. Lin, and L. Zhou. 2005. A BE-based Multi-document Summarization with Sentence Compression. In *Proceedings of Multilingual Summarization Evaluation* (ACL 2005 workshop), Ann Arbor, MI.

Knight, K., and Marcu, D. 2000. Statistics-Based Summarization: Step One: Sentence Compression. *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence AAAI 2000*, Austin, Texas, July 30-Augeust 3, 2000.

Lin, C-Y. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on the Text Summarization Branches Out* (WAS 2004), Barcelona, Spain, July 25-26, 2004.

Lin, C-Y. and E. Hovy. 1997. Identify Topics by Position. *Proceedings of the 5th Conference on Applied Natural Language Processing* (ANLP), Washington, D.C.

Lin, C-Y, and E. H. Hovy. 2000. The Automtated Acquisition of Topic Signatures for Text Summarization. *Proceedings of the COLING Conference*, Strasbourg, France. August 2000.

Lin, C.-Y. and E. Hovy. 2002. Automated Multi-document summarization in NeATS. *Proceedings of the Human Language Technology Conference* (HLT2002), San Diego, CA, USA, March 23-27, 2002.

Mani, I., Gates, B., and Bloedorn, E. 1999. Improving Summaries by Revising Them. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, 558-565.

Nenkova, A. and R. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. *Proceedings of the HLT-NAACL conference.* Boston, MA.

Van Halteren, H. and S. Teufel. 2003. Examining the Consensus between Human Summaries: Initial Experiments with Factoid Analysis. *Proceedings of the HLT-NAACL Workshop on Automatic Summarization*. Edmonton, Canada.