NEATS: A Multidocument Summarizer

Chin-Yew Lin and Eduard Hovy

USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{cyl,hovy}@isi.edu
tel: 310-448-8711 and 310-448-8731

August 2001

1. System

NEATS is a multidocument summarization system that attempts to extract relevant or interesting portions from a set of documents about some topic and present them in coherent order. It is tailored to the genre of newspaper news articles, and it works for English, but can be made multilingual without a great deal of effort. At present NEATS produces generic (author's point of view) summaries, but it could be made sensitive to desired focus topics, input by a user.

Given an input of a collection of sets of newspaper articles, NEATS applies the following 6 steps.

1. Extract and rank passages

Given the input documents, form a query, extract sentences, and rank them, using modules of the Webclopedia QA system (Hovy et al., 2000):

1.a identify key concepts for each topic group. Compute unigram, bigram, and trigram topic signatures (Lin and Hovy, 2000; Hovy and Lin, 1999) for each group, using the likelihood ratio λ (Dunning, 1993)

1.b remove from the signatures all words or phrases that occur in fewer than half the texts of the topic group

1.c save the signatures in a tree, organized by signature overlap, using the parse tree format of CONTEX (Hermjakob, 1997; 2000); see Figure 1

1.d use the Webclopedia query formation module to form queries, most specific first

1.e use Webclopedia's version of MG to perform sentence-level IR and return a ranked list of sentences (Hovy et al., 2000).

2. Filter for content

Using an OPP policy as developed for the SUMMARIST single-document summarizer (Lin and Hovy, 1997), remove extracted sentences too far from the high-importance regions:

2.a from the ranked list, remove all sentences with sentence position > 10 (simple OPP policy)

2.b also decrease ranking score of all sentence containing stigma words (day names; time expressions; sentences starting with conjunctions such as "but", "although"; sentences containing quotation marks; sentences containing the verb "say").

3. Enforce cohesion and coherence

Each remaining sentence is paired with a suitable introductory sentence:

3.a pair each sentence with the first sentence (lead) of its document; but if the first sentence contains fewer than 5 words, then take the next one. For example (where *x.y* stands for *document number* . *sentence number*):

4.3, 6.6, 2.5, 5.2… $\rightarrow$ 4.1, 4.3, 6.1, 6.6, 2.1, 2.5, 5.1, 5.2...

4. Filter for length

Select the required number of sentence pairs using a simplified version of CMU's MMR algorithm:
4.a include first pair
4.b using a simplified version of MMR (Goldstein et al., 1999), find the sentence pair most different from the included ones, and include it too. (In the DUC-2001 implementation, NEATS did not consider the sentence pair, just the sentence. This caused some degradation.)
4.c repeat 4.b until the summary length criterion is satisfied

$\rightarrow$ 4.1, 4.3, 2.1, 2.5

5. Ensure chronological coherence

Reorder the pairs in publication order, and disambiguate all time words with explicit dates:
5.a reorder pairs in publication order

$\rightarrow$ 2.1, 2.5, 4.1, 4.5

5.b for each time word ("today", "Monday", etc.) compute the actual date (from the dateline) and include it in the text in parentheses, in order to signal which day each "today" (etc.) is.

6. Format and printing results
Format and output the final result.

2. Discussion

This simple algorithm gives surprisingly reasonable results. We like the following aspects.

Typical current extractive summarization methods are essentially IR in miniature: from a set of sentences (instead of texts), select and rank the ones most relevant to the query. The major problems are **creating the query** and then **assembling the extracted sentences into a single coherent text** (a step that IR does not have).

For creating the query, we saved a great deal of development time by using existing modules from SUMMARIST (Hovy and Lin, 1999) and Webclopedia (Hovy et al., 2000; 2001). SUMMARIST's topic signature creation techniques (Lin and Hovy, 2000) allowed us directly to compute a ranked list of words (and bi- and trigrams) most characteristic of each document set. By placing these ngrams (and their sub-ngrams, which form a cluster) into the parse tree format we use for the IR stage of Webclopedia (Figure 1), we could directly form increasingly general queries, extract the most relevant sentences from the document set, and rank them.

```
(:SURF "WEBCL-SUMMMARIZER-HOSPITAL"
   :CAT S-NP
   :CLASS I-EN-WEBCL-SIGNATURE-HOSPITAL
   :LEX  0.9
   :SUBS (((HOSPITAL-0)
               (:SURF "Hospital Health Center"
               :CAT S-NP
               :CLASS I-EN-WEBCL-SIGNATURE-HOSPITAL
               :LEX 0.6
               :SUBS (((HOSPITAL-14)
                           (:SURF "Hospital Health"
                           :CAT S-NP
                           :CLASS I-EN-WEBCL-SIGNATURE-HOSPITAL
                           :LEX 0.6))
                      ((HOSPITAL-24)
                           (:SURF "Health Center"
                           :CAT S-NP
                           :CLASS I-EN-WEBCL-SIGNATURE-HOSPITAL
                           :LEX 0.6))
                      ((HOSPITAL-37) (:SURF "Hospital" …))
                      ((HOSPITAL-43) (:SURF "Center" …))
```

((HOSPITAL-44) (:SURF “Health” …)))))
((HOSPITAL-25) (:SURF “John Hospital” …)))
((HOSPITAL-42) (:SURF “doctors”…))
((HOSPITAL-45) (:SURF “Marina” …))
((HOSPITAL-47) (:SURF “April” …))
((HOSPITAL-49) (:SURF “sinus” …))))
…

Figure 1. Portion of topic signature cluster tree for “Hospital Health Center” ngrams.

To assemble the extracted sentences into a single coherent text, we used the fact that a lead sentence, which introduces the article, is a powerful context-setter for each nearby (early) sentence in the article. We therefore paired each extracted sentence with its lead sentence, selected as appropriate.

One further factor interfering with coherence was misleading time words: “today” in articles written on different days means different dates. To disambiguate all time words we therefore computed the actual dates from the articles’ datelines and included them after each time word. A typical summary is included in Figure 2.

> (07/19/89) Simply put, the question was who should be counted as a person and who, if anybody, should not.
>
> (07/19/89) The point at issue in Senate debate on a new immigration bill was whether illegal aliens should be counted in the process that will reallocate House seats among states after the 1990 census.
>
> (09/30/89) In a blow to California and other states with large immigrant populations, the Senate voted Friday (09/29/89) to bar the Census Bureau from counting illegal aliens in the 1990 population count.
>
> (09/30/89) At stake are the number of seats in Congress for California, Florida, New York, Illinois, Pennsylvania and other states that will be reapportioned on the basis of next year's census.

Figure 2. Example 100-word summary.

3. Results

We were surprised by the content and readability of the results. Analyzing all systems’ results for DUC-2001, we made the following observations.

1. With respect to **content**, we computed Recall, Precision, and F-Measure using the following formulas:

Recall = (# of model units marked with peer units) / (# of model units)
Precision = (# of unique peer units marked with model units) / (# of peer units)
F-Measure = 2*Recall*Precision / (Recall+Precision)

According to this, NEATS did not fare badly (though its relative rank may change with different definitions of Recall and Precision). Systems’ scores using these formulas are shown in the histogram in Figure 3. Humans did better than any system (both humans over 55%), outscoring the nearest system by about 10%. Only 1 system (NEATS) scored in the 40s, with 45%. 6 systems scored between 35% and 40%, and 3 scored between 30% and 35%. Despite the low inter-human agreement (which we take to reflect the undefinedness of the ‘generic summary’ task), there is obviously still considerable room for systems to improve. We expect that systems that compress their output (unlike NEATS) will thereby gain more space to include additional important material.

2. When it came to the measures for **grammar, coherence, and cohesion**, the results are confusing. If even the human-made summaries score only 3.8 / 4 for grammaticality, 2.63 / 4 for cohesion, and 3.14 / 4 for coherence, it is unclear what these categories mean, or how the assessors arrived at these scores. Grammaticality, surely, should be easy to judge for sentences such as these. NEATS, being an extraction system, delivers pure newspaper sentences (with

dates added in parentheses); its grammaticality should have been 100%, assuming the newspaper journalists were competent. Yet it scored only 3.72 / 4 for grammaticality. In fact, we would guess that only one system (Y), whose grammaticality score is much lower than any other systems', tried to do something interesting with sentence structure; all the others are probably pure extraction systems.

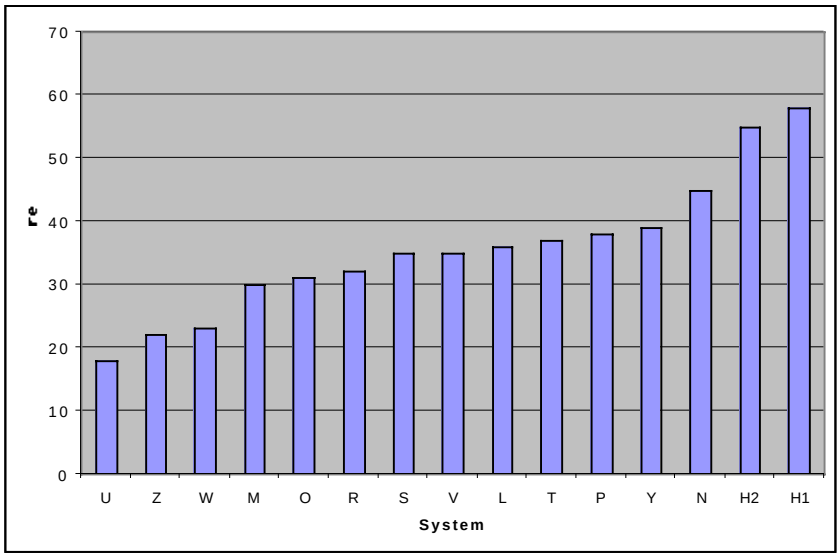

Figure 3. Histogram of F-Measures.

4. References

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

Goldstein, J., M. Kantrowitz, V. Mittal, and J. Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, Berkeley, CA, 121–128.

Hermjakob, U. 1997. *Learning Parse and Translation Decisions from Examples with Rich Context.* Ph.D. dissertation, University of Texas at Austin. file://ftp.cs.utexas.edu/pub/~mooney/papers/hermjakob-dissertation-97.ps.gz.

Hermjakob, U. 2000. Rapid Parser Development: A Machine Learning Approach for Korean. In *Proceedings of the North American chapter of the Association for Computational Linguistics* (NAACL-2000). http://www.isi.edu/~ulf/papers/kor_naacl00.ps.gz.

Hovy, E.H. and C.-Y. Lin. 1999. Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds), *Advances in Automatic Text Summarization*. Cambridge: MIT Press.

Hovy, E.H., L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. 2000. Question Answering in Webclopedia. *Proceedings of the TREC-9 Conference.* NIST, Gaithersburg, MD. November 2000.

Hovy, E.H., L. Gerber, U. Hermjakob, C.-Y. Lin, and D. Ravichandran. 2001. Toward Semantics-Based Answer Pinpointing. *Proceedings of the DARPA Human Language Technology Conference (HLT).* San Diego, CA. March 2001.

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

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