In Proceedings of the HLT-NAACL 2003 Workshop on Automatic Summarization, May 30 to June 1, 2003, Edmonton, Canada.

The Potential and Limitations of Automatic Sentence Extraction for Summarization

Chin-Yew Lin and Eduard Hovy
University of Southern California/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
{cyl,hovy}@isi.edu

Abstract

In this paper we present an empirical study of the potential and limitation of sentence extraction in text summarization. Our results show that the single document generic summarization task as defined in DUC 2001 needs to be carefully refocused as reflected in the low inter-human agreement at 100-word [1] (0.40 score) and high upper bound at full text [2] (0.88) summaries. For 100-word summaries, the performance upper bound, 0.65, achieved oracle extracts[3]. Such oracle extracts show the promise of sentence extraction algorithms; however, we first need to raise inter-human agreement to be able to achieve this performance level. We show that compression is a promising direction and that the compression ratio of summaries affects average human and system performance.

1 Introduction

Most automatic text summarization systems existing today are extraction systems that extract parts of original documents and output the results as summaries. Among them, sentence extraction is by far the most popular (Edmundson 1969, Luhn 1969, Kupiec et al. 1995, Goldstein et al. 1999, Hovy and Lin 1999). The majority of systems participating in the past Document Understanding Conference (DUC 2002), a large scale summarization evaluation effort sponsored by the US government, are extraction based. Although systems based on information extraction (Radev and McKeown 1998, White et al. 2001, McKeown et al. 2002) and discourse analysis (Marcu 1999b, Strzalkowski et al. 1999) also exist, we focus our study on the potential and limitations of sentence extraction systems with the hope that our results will further progress in most of the automatic text summarization systems and evaluation setup.

The evaluation results of the single document summarization task in DUC 2001 and 2002 (DUC 2002, Paul & Liggett 2002) indicate that most systems are as good as the baseline lead-based system and that humans are significantly better, though not by much. This leads to the belief that lead-based summaries are as good as we can get for single document summarization in the news genre, implying that the research community should invest future efforts in other areas. In fact, a very short summary of about 10 words (headline-like) task has replaced the single document 100-word summary task in DUC 2003. The goal of this study is to renew interest in sentence extraction-based summarization and its evaluation by estimating the performance upper bound using oracle extracts, and to highlight the importance of taking into account the compression ratio when we evaluate extracts or summaries.

Section 2 gives an overview of DUC relevant to this study. Section 3 introduces a recall-based unigram co-occurrence automatic evaluation metric. Section 4 presents the experimental design. Section 5 shows the empirical results. Section 6 concludes this paper and discusses future directions.

[1] We compute unigram co-occurrence score of a pair of manual summaries, one as candidate summary and the other as reference.

[2] We compute unigram co-occurrence scores of a full text and its manual summaries of 100 words. These scores are the best achievable using the unigram co-occurrence scoring metric since all possible words are contained in the full text. Three manual summaries are used.

[3] Oracle extracts are the best scoring extracts generated by exhaustive search of all possible sentence combinations of 100±5 words.

2 Document Understanding Conference

Fully automatic single-document summarization was one of two main tasks in the 2001 Document Understanding Conference. Participants were required to create a generic 100-word summary. There were 30 test sets in DUC 2001 and each test set contained about 10 documents. For each document, one summary was created manually as the ëidealí model summary at approximately 100 words. We will refer to this manual summary as H1. Two other manual summaries were also created at about that length. We will refer to these two additional human summaries as H2 and H3. In addition, baseline summaries were created automatically by taking the first n sentences up to 100 words. We will refer this baseline extract as B1.

3 Unigram Co-Occurrence Metric

In a recent study (Lin and Hovy 2003), we showed that the recall-based unigram co-occurrence automatic scoring metric correlated highly with human evaluation and has high recall and precision in predicting statistical significance of results comparing with its human counterpart. The idea is to measure the content similarity between a system extract and a manual summary using simple n-gram overlap. A similar idea called IBM BLEU score has proved successful in automatic machine translation evaluation (Papineni et al. 2001, NIST 2002). For summarization, we can express the degree of content overlap in terms of n-gram matches as the following equation:

$$C_n = \frac{\sum_{C \in \{Model\ Units\}} \sum_{n-gram \in C} Count_{match}(n-gram)}{\sum_{C \in \{Model\ Units\}} \sum_{n-gram \in C} Count(n-gram)} \quad (1)$$

Model units are segments of manual summaries. They are typically either sentences or elementary discourse units as defined by Marcu (1999b). $Count_{match}$(*n-gram*) is the maximum number of *n-gram*s co-occurring in a system extract and a model unit. *Count*(*n-gram*) is the number of *n-gram*s in the model unit. Notice that the average n-gram coverage score, C_n, as shown in equation 1, is a recall-based metric, since the denominator of equation 1 is the sum total of the number of *n-gram*s occurring in the model summary instead of the system summary and only one model summary is used for each evaluation. In summary, the unigram co-occurrence statistics we use in the following sections are based on the following formula:

$$Ngram(i, j) = \exp\left(\sum_{n=i}^{j} w_n \log C_n\right) \quad (2)$$

Where $j \geq i$, i and j range from 1 to 4, and w_n is $1/(j-i+1)$. *Ngram*(1, 4) is a weighted variable length *n*-gram match score similar to the IBM BLEU score; while *Ngram*(*k*, *k*), i.e. $i = j = k$, is simply the average *k*-gram co-occurrence score C_k. In this study, we set $i = j = 1$, i.e. unigram co-occurrence score.

With a test collection available and an automatic scoring metric defined, we describe the experimental setup in the next section.

4 Experimental Designs

As stated in the introduction, we aim to find the performance upper bound of a sentence extraction system and the effect of compression ratio on its performance. We present our experimental designs to address these questions in the following sections.

4.1 Performance Upper Bound Estimation Using Oracle Extract

In order to estimate the potential of sentence extraction systems, it is important to know the upper bound that an ideal sentence extraction method might achieve and how far the state-of-the-art systems are away from the bound. If the upper bound is close to state-of-the-art systemsí performance then we need to look for other summarization methods to improve performance. If the upper bound is much higher than any current systems can achieve, then it is reasonable to invest more effort in sentence extraction methods. The question is how to estimate the performance upper bound. Our solution is to cast this estimation problem as an optimization problem. We exhaustively generate all possible sentence combinations that satisfy given length constraints for a summary, for example, all the sentence combinations totaling 100±5 words. We then compute the unigram co-occurrence score for each sentence combination, against the ideal. The best combinations are the ones with the highest unigram co-occurrence score. We call this sentence combination the *oracle* extract. Figure 1 shows an oracle extract for document AP900424-0035. One of its human summaries is shown in Figure 2. The oracle extract covers almost all aspects of the human summary except sentences 5 and 6 and part of sentence 4. However, if we allow the automatic extract to contain more words, for example, 150 words shown in Figure 3, the longer oracle extract then covers everything in the human summary. This indicates that lower compression can boost system performance. The ultimate effect of compression can be computed using the full text as the oracle extract, since the full text should contain everything included in the human summary. That situation provides the best achievable unigram co-occurrence score. A near optimal score also confirms the validity of using the unigram co-occurrence scoring method as an automatic evaluation method.

```
<DOC>
<DOCNO>AP900424-0035</DOCNO>
<DATE>04/24/90</DATE>
<HEADLINE>
<S HSNTNO="1">Elizabeth Taylor in Intensive Care Unit</S>
<S HSNTNO="2">By JEFF WILSON</S>
<S HSNTNO="3">Associated Press Writer</S>
<S HSNTNO="4">SANTA MONICA, Calif. (AP)</S>
</HEADLINE>
<TEXT>
<S SNTNO="1">A seriously ill Elizabeth Taylor battled pneumonia at her
hospital, her breathing assisted by a ventilator, doctors say.</S>
<S SNTNO="2">Hospital officials described her condition late Monday
as stabilizing after a lung biopsy to determine the cause of the pneumo-
nia.</S>
<S SNTNO="3">Analysis of the tissue sample was expected to take until
Thursday, said her spokeswoman, Chen Sam.</S>
<S SNTNO="9">Another spokewoman for the actress, Lisa Del Favaro,
said Miss Taylor's family was at her bedside.</S>
<S SNTNO="13">``It is serious, but they are really pleased with her
progress.</S>
<S SNTNO="22">During a nearly fatal bout with pneumonia in 1961,
Miss Taylor underwent a tracheotomy, an incision into her windpipe to
help her breathe.</S>
</TEXT>
</DOC>
```

Figure 1. A 100-word oracle extract for document AP900424-0035.

```
<DOC>
<TEXT>
<S SNTNO="1">Elizabeth Taylor battled pneumonia at her hospital,
assisted by a ventilator, doctors say.</S>
<S SNTNO="2">Hospital officials described her condition late Monday
as stabilizing after a lung biopsy to determine the cause of the pneumo-
nia.</S>
<S SNTNO="3">Analysis of the tissue sample was expected to be com-
plete by Thursday.</S>
<S SNTNO="4">Ms. Sam, spokeswoman said "it is serious, but they are
really pleased with her progress.</S>
<S SNTNO="5">She's not well.</S>
<S SNTNO="6">She's not on her deathbed or anything.</S>
<S SNTNO="7">Another spokeswoman, Lisa Del Favaro, said Miss
Taylor's family was at her bedside.</S>
<S SNTNO="8">During a nearly fatal bout with pneumonia in 1961, Miss
Taylor underwent a tracheotomy to help her breathe.</S>
</TEXT>
</DOC>
```

Figure 2. A manual summary for document AP900424-0035.

4.2 Compression Ratio and Its Effect on System Performance

One important factor that affects the average performance of sentence extraction system is the number of sentences contained in the original documents. This factor is often overlooked and has never been addressed systematically. For example, if a document contains only one sentence then this document will not be useful in differentiating summarization system performance ñ there is only one choice. However, for a document of 100 sentences and assuming each sentence is 20 words long, there are $C(100,5)$ = 75,287,520 different 100-word extracts. This huge search space lowers the chance of agreement between humans on what constitutes a good summary. It also makes system and human performance approach *average* since it is more likely to include some good sentences but not all of them. Empirical results shown in Section 5 confirm this and that leads us to the question of how to construct a corpus to evaluate summarization systems. We discuss this issue in the conclusion section.

```
<DOC>
<DOCNO>AP900424-0035</DOCNO>
<DATE>04/24/90</DATE>
<HEADLINE>
<S HSNTNO="1">Elizabeth Taylor in Intensive Care Unit</S>
<S HSNTNO="2">By JEFF WILSON</S>
<S HSNTNO="3">Associated Press Writer</S>
<S HSNTNO="4">SANTA MONICA, Calif. (AP)</S>
</HEADLINE>
<TEXT>
<S SNTNO="1">A seriously ill Elizabeth Taylor battled pneumonia at her
hospital, her breathing assisted by a ventilator, doctors say.</S>
<S SNTNO="2">Hospital officials described her condition late Monday
as stabilizing after a lung biopsy to determine the cause of the pneumo-
nia.</S>
<S SNTNO="3">Analysis of the tissue sample was expected to take until
Thursday, said her spokeswoman, Chen Sam.</S>
<S SNTNO="4">The 58-year-old actress, who won best-actress Oscars
for ``Butterfield 8" and ``Who's Afraid of Virginia Woolf," has been
hospitalized more than two weeks.</S>
<S SNTNO="8">Her condition is presently stabilizing and her physicians
are pleased with her progress."</S>
<S SNTNO="9">Another spokewoman for the actress, Lisa Del Favaro,
said Miss Taylor's family was at her bedside.</S>
<S SNTNO="13">``It is serious, but they are really pleased with her
progress.</S>
<S SNTNO="14">She's not well.</S>
<S SNTNO="15">She's not on her deathbed or anything," Ms. Sam said
late Monday.</S>
<S SNTNO="22">During a nearly fatal bout with pneumonia in 1961,
Miss Taylor underwent a tracheotomy, an incision into her windpipe to
help her breathe.</S>
</TEXT>
</DOC>
```

Figure 3. A 150-word oracle extract for document AP900424-0035.

4.3 Inter-Human Agreement and Its Effect on System Performance

In this section we study how inter-human agreement affects system performance. Lin and Hovy (2002) reported that, compared to a manually created ideal, humans scored about 0.40 in average coverage score and the best system scored about 0.35. According to these numbers, we might assume that humans cannot agree to each other on what is important and the best system is almost as good as humans. If this is true then estimating an upper bound using oracle extracts is meaningless. No matter how high the estimated upper bounds may be, we probably would never be able to achieve that performance due to lack of agreement between humans: the oracle approximating one human would fail miserably with another. Therefore we set up experiments to investigate the following:

1. What is the distribution of inter-human agreement?

2. How does a state-of-the-art system differ from average human performance at different inter-human agreement levels?

We present our results in the next section using 303 newspaper articles from the DUC 2001 single document summarization task. Besides the original documents, we also have three human summaries, one lead summary (B1), and one automatic summary from one top performing system (T) for each document.

5 Results

In order to determine the empirical upper and lower bounds of inter-human agreement, we first ran cross-human evaluation using unigram co-occurrence scoring through six human summary pairs, i.e. (H1,H2), (H1,H3), (H2,H1), (H2,H3), (H3,H1), and (H3,H2). For a summary pair (X,Y), we used X as the model summary and Y as the system summary. Figure 4 shows the distributions of four different scenarios. The MaxH distribution picks the best inter-human agreement scores for each document, the MinH distribution the minimum one, the MedH distribution the median, and the AvgH distribution the average. The average of the best inter-human agreement and the average of average inter-human agreement differ by about 10 percent in unigram co-occurrence score and 18 percent between MaxH and MinH. These big differences might come from two sources. The first one is the limitation of the unigram

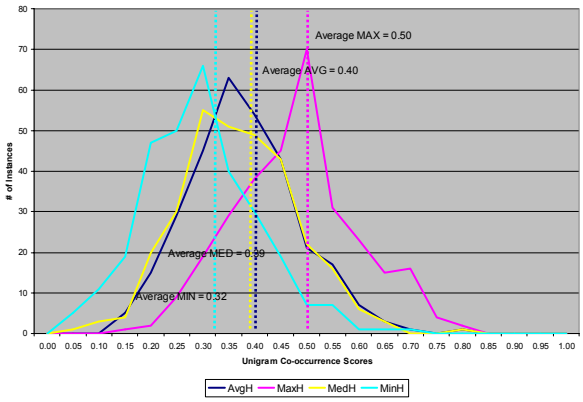

Figure 4. DUC 2001 single document inter-human unigram co-occurrence score distributions for maximum, minimum, average, and median.

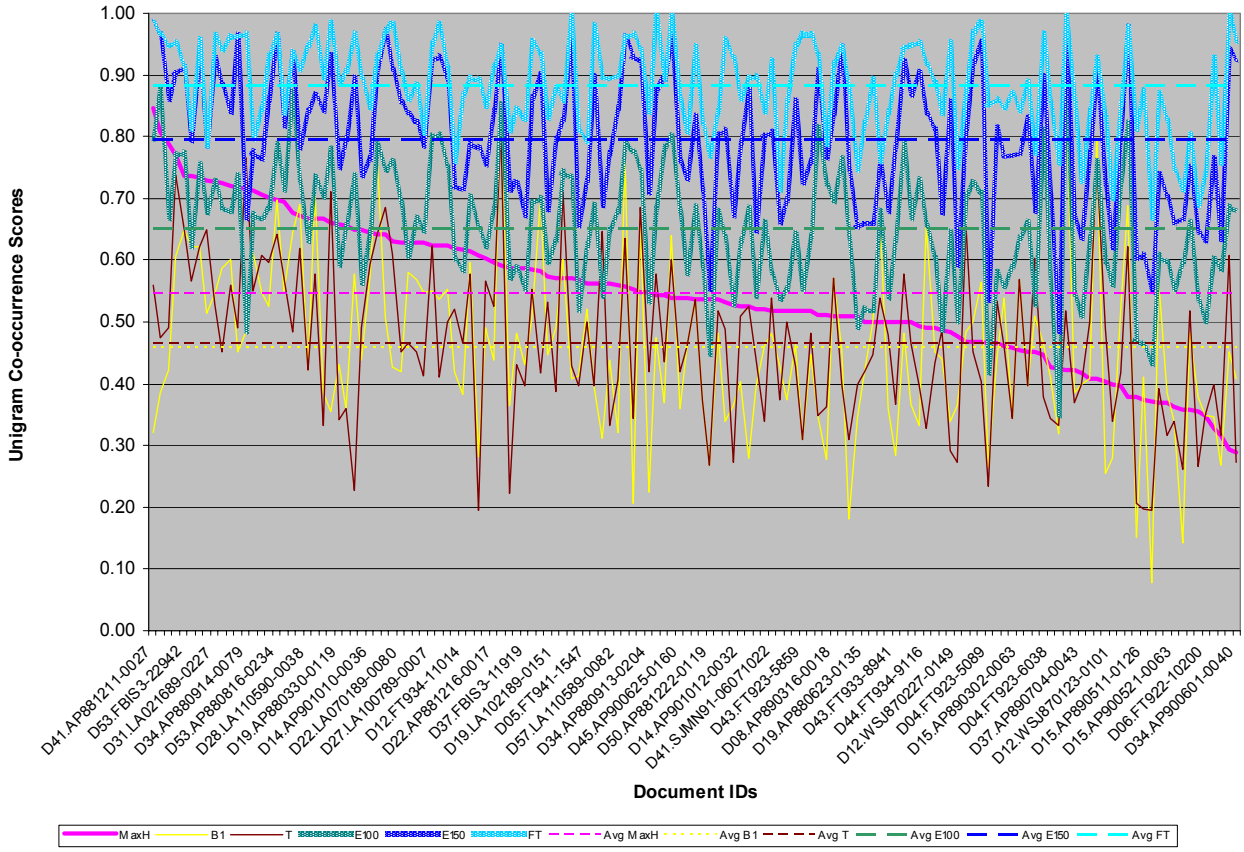

Figure 5. DUC 2001 single document inter-human, baseline, system, 100-word, 150-word, and full text oracle extracts unigram co-occurrence score distributions (# of sentences<=30). Document IDs are sorted by decreasing MaxH.

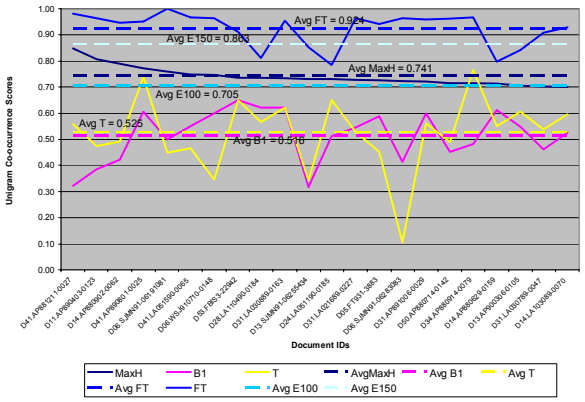

Figure 6. DUC 2001 single document inter-human, baseline, system, and full text unigram co-occurrence score distributions (Set A).

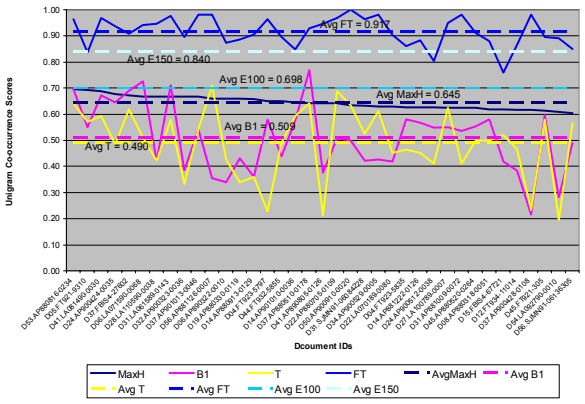

Figure 7. DUC 2001 single document inter-human, baseline, system, and full text unigram co-occurrence score distributions (Set B).

co-occurrence scoring applied to manual summaries that it cannot recognize synonyms or paraphrases. The second one is the true lack of agreement between humans. We would like to conduct an in-depth study to address this question, and would just assume the unigram co-occurrence scoring is reliable.

In other experiments, we used the best inter-human agreement results as the reference point for human performance upper bound. This also implied that we used the human summary achieving the best inter-human agreement score as our reference summary.

Figure 5 shows the unigram co-occurrence scores of human, baseline, system T, and three oracle extraction systems at different extraction lengths. We generated all possible sentence combinations that satisfied 100±5 words constraints. Due to computation-intensive nature of this task, we only used documents with fewer than 30 sentences. We then computed the unigram co-occurrence score for each combination, selected the best one as the oracle extraction, and plotted the score in the

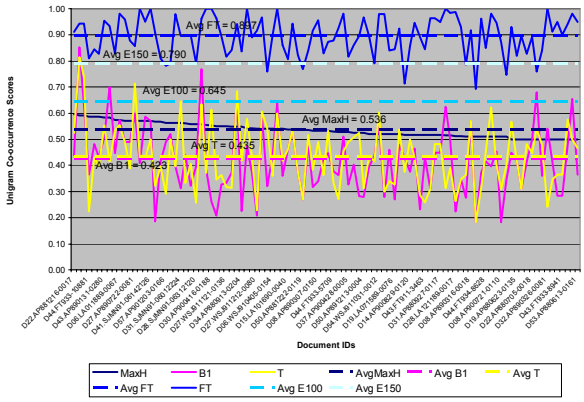

Figure 8. DUC 2001 single document inter-human, baseline, system, and full text unigram co-occurrence score distributions (Set C).

figure. The curve for 100±5 words oracle extractions is the upper bound that a sentence extraction system can achieve within the given word limit. If an automatic system is allowed to extract more words, we can expect that longer extracts would boost system performance. The question is how much better and what is the ultimate limit? To address these questions, we also computed unigram co-occurrence scores for oracle extractions of 150±5 words and full text[4]. The performance of full text is the ultimate performance an extraction system can reach using the unigram co-occurrence scoring method. We also computed the scores of the lead baseline system (B1) and an automatic system (T).

The average unigram co-occurrence score for full text (FT) was 0.833, 150±5 words (E150) was 0.796, 100±5 words (E100) was 0.650, the best inter-human agreement (MaxH) was 0.546, system T was 0.465, and baseline was 0.456. It is interesting to note that the state-of-the-art system performed at the same level as the baseline system but was still about 10% away from human. The 10% difference between E100 and MaxH (0.650 vs. 0.546) implies we might need to constraint humans to focus their summaries in certain aspects to boost inter-human agreement to the level of E100; while the 15% and 24% improvements from E100 to E150 and FT indicate compression would help push overall system performance to a much higher level, if a system is able to compress longer summaries into a shorter without losing important content.

To investigate relative performance of humans, systems, and oracle extracts at different inter-human agreement levels, we created three separate document sets based on their maximum inter-human agreement (MaxH) scores. Set Set A had MaxH score greater than or equal to 0.70, set B was between 0.70 and 0.60, and

[4] We used full text as extract and computed its unigram co-occurrence score against a reference summary.

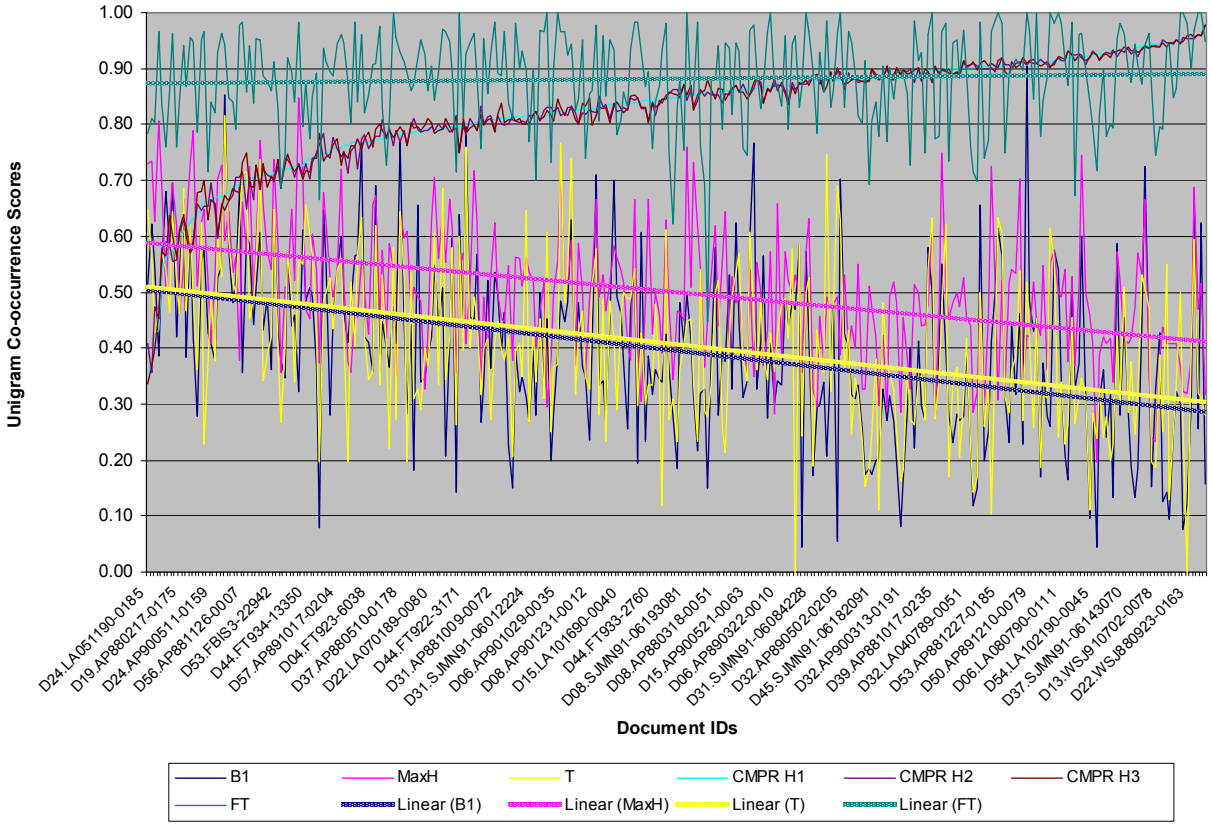

Figure 9. DUC 2001 single doc inter-human, baseline, and system unigram co-occurrence score versus compression ratio. Document IDs are sorted by increasing compression ratio CMPR H1.

set C between 0.60 and 0.50. A had 22 documents, set B 37, and set C 100. Total was about 52% (=159/303) of the test collection. The 100±5 and 150±5 words averages were computed over documents which contain at most 30 sentences. The results are shown in Figures 6, 7, and 8. In the highest inter-human agreement set (A), we found that average MaxH, 0.741, was higher than average 100±5 words oracle extract, 0.705; while the average automatic system performance was around 0.525. This is good news since the high inter-human agreement and the big difference (0.18) between 100±5 words oracle and automatic system performance presents a research opportunity for improving sentence extraction algorithms. The scores of MaxH (0.645 for set B and 0.536 for set C) in the other two sets are both lower than 100±5 words oracles (0.698 for set B, 5.3% lower, and 0.645 for set C, 9.9% lower). This result suggests that optimizing sentence extraction algorithms at the Set C level might not be worthwhile since the algorithms are likely to overfit the training data. The reason is that the average run time performance of a sentence extraction algorithm depends on the maximum inter-human agreement. For example, given a training reference summary T_{SUM1} and its full document T_{DOC1}, we optimize our sentence extraction algorithm to generate an oracle extract based on T_{SUM1} from T_{DOC1}. In the run time, we test on a reference summary R_{SUM1} and its full document R_{DOC1}. In the unlikely case that R_{DOC1} is the same as T_{DOC1} and R_{SUM1} is the same as T_{SUM1}, i.e. T_{SUM1} and R_{SUM1} have unigram co-occurrence score of 1 (perfect inter-human agreement for two summaries of one document), the optimized algorithm will generate a perfect extract for R_{DOC1} and achieve the best performance since it is optimized on T_{SUM1}. However, usually T_{SUM1} and R_{SUM1} are different. Then the performance of the algorithm will not exceed the maximum unigram co-occurrence score between T_{SUM1} and R_{SUM1}. Therefore it is important to ensure high inter-human agreement to allow researchers room to optimize sentence extraction algorithms using oracle extracts.

Finally, we present the effect of compression ratio on inter-human agreement (MaxH) and performance of baseline (B1), automatic system T (T), and full text oracle (FT) in Figure 9. Compression ratio is computed in terms of words instead of sentences. For example, a 100 words summary of a 500 words document has a compression ratio of 0.80 (=1ñ100/500). The figure shows that three human summaries (H1, H2, and H3) had different compression ratios (CMPR H1, CMPR H2, and CMPR H3) for different documents but did not differ

much. The unigram co-occurrence scores for B1, T, and MaxH were noisy but had a general trend (Linear B1, Linear T, and Linear MaxH) of drifting into lower performance when compression ratio increased (i.e. when summaries became shorter); while the performance of FT did not exhibit a similar trend. This confirms our earlier hypothesis that humans are less likely to agree at high compression ratio and system performance will also suffer at high compression ratio. The constancy of FT across different compression ratios is reasonable since FT scores should only depend on how well the unigram co-occurrence scoring method captures content overlap between a full text and its reference summaries and how likely humans use vocabulary outside the original document.

6 Conclusions

In this paper we presented an empirical study of the potential and limitations of sentence extraction as a method of automatic text summarization. We showed the following:

(1) How to use oracle extracts to estimate the performance upper bound of sentence extraction methods at different extract lengths. We understand that summaries optimized using unigram co-occurrence score do not guarantee good quality in terms of coherence, cohesion, and overall organization. However, we would argue that a good summary does require good content and we will leave how to make the content cohesive, coherent, and organized to future research.

(2) Inter-human agreement varied a lot and the difference between maximum agreement (MaxH) and minimum agreement (MinH) was about 18% on the DUC 2001 data. To minimize the gap, we need to define the summarization task better. This has been addressed by providing guided summarization tasks in DUC 2003 (DUC 2002). We guesstimate the gap should be smaller in DUC 2003 data.

(3) State-of-the-art systems performed at the same level as the baseline system but were still about 10% away from the average human performance.

(4) The potential performance gains (15% from E100 to E150 and 24% to FT) estimated by oracle extracts of different sizes indicated that sentence compression or sub-sentence extraction are promising future directions.

(5) The relative performance of humans and oracle extracts at three inter-human agreement intervals showed that it was only meaningful to optimize sentence extraction algorithms if inter-human agreement was high. Although overall high inter-human agreement was low but subsets of high inter-human agreement did exist. For example, about human achieved at least 60% agreement in 59 out of 303 (~19%) documents of 30 sentences or less.

(6) We also studied how compression ratio affected inter-human agreement and system performance, and the results supported our hypothesis that humans tend to agree less at high compression ratio, and similar between humans and systems. How to take into account this factor in future summarization evaluations is an interesting topic to pursue further.

Using exhaustive search to identify oracle extraction has been studied by other researchers but in different contexts. Marcu (1999a) suggested using exhaustive search to create training extracts from abstracts. Donaway et al. (2000) used exhaustive search to generate all three sentences extracts to evaluate different evaluation metrics. The main difference between their work and ours is that we searched for extracts of a fixed number of words while they looked for extracts of a fixed number of sentences.

In the future, we would like to apply a similar methodology to different text units, for example, sub-sentence units such as elementary discourse unit (Marcu 1999b). We want to study how to constrain the summarization task to achieve higher inter-human agreement, train sentence extraction algorithms using oracle extracts at different compression sizes, and explore compression techniques to go beyond simple sentence extraction.

References

Donaway, R.L., Drummey, K.W., and Mather, L.A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In Proceeding of the Workshop on Automatic Summarization, post-conference workshop of ANLP-NAACL-2000, Seattle, WA, USA, 69ñ78.

DUC. 2002. The Document Understanding Conference. http://duc.nist.gov.

Edmundson, H.P. 1969. New Methods in Automatic Abstracting. Journal of the Association for Computing Machinery. 16(2).

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, USA, 121ñ128.

Hovy, E. and C.-Y. Lin. 1999. Automatic Text Summarization in SUMMARIST. In I. Mani and M.

Maybury (eds), Advances in Automatic Text Summarization, 81ñ94. MIT Press.

Kupiec, J., J. Pederson, and F. Chen. 1995. A Trainable Document Summarizer. In Proceedings of the 18th International ACM Conference on Research and Development in Information Retrieval (SIGIR-95), Seattle, WA, USA, 68ñ73.

Lin, C.-Y. and E. Hovy. 2002. Manual and Automatic Evaluations of Summaries. In Proceedings of the Workshop on Automatic Summarization, post-conference workshop of ACL-2002, pp. 45-51, Philadelphia, PA, 2002.

Lin, C.-Y. and E.H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada, May 27 ñ June 1, 2003.

Luhn, H. P. 1969. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 2(2), 1969.

Marcu, D. 1999a. The automatic construction of large-scale corpora for summarization research. Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley, CA, USA, 137ñ144.

Marcu, D. 1999b. Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds), Advances in Automatic Text Summarization, 123ñ 136. MIT Press.

McKeown, K., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, S. Sigelman. 2002. Tracking and Summarizing News on a Daily Basis with Columbiaís Newsblaster. In Proceedings of Human Language Technology Conference 2002 (HLT 2002). San Diego, CA, USA.

NIST. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics.

Over, P. and W. Liggett. 2002. Introduction to DUC-2002: an Intrinsic Evaluation of Generic News Text Summarization Systems. *In Proceedings of Workshop on Automatic Summarization* (DUC 2002), Philadelphia, PA, USA.
http://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022).

Radev, D.R. and K.R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-line Sources. Computational Linguistics, 24(3):469ñ500.

Strzalkowski, T, G. Stein, J. Wang, and B, Wise. A Robust Practical Text Summarizer. 1999. In I. Mani and M. Maybury (eds), Advances in Automatic Text Summarization, 137ñ154. MIT Press.

White, M., T. Korelsky, C. Cardie, V. Ng, D. Pierce, and K. Wagstaff. 2001. Multidocument Summarization via Information Extraction. In Proceedings of Human Language Technology Conference 2001 (HLT 2001), San Diego, CA, USA.