A Sentence-Trimming Approach to Multi-Document Summarization

David Zajic, Bonnie Dorr, Jimmy Lin, Christof Monz
Department of Computer Science
University of Maryland
College Park, MD 20742
{dmzajic,bonnie,jimmylin,christof}
@umiacs.umd.edu

Richard Schwartz
BBN Technologies
9861 Broken Land Parkway, Suite 156
Columbia, MD 21046
schwartz@bbn.com

Abstract

We implemented an initial application of a sentence-trimming approach (Trimmer) to the problem of multi-document summarization in the MSE2005 and DUC2005 tasks. Sentence trimming was incorporated into a feature-based summarization system, called Multi-Document Trimmer (MDT), by using sentence trimming as both a pre-processing stage and a feature for sentence ranking. We demonstrate that we were able to port Trimmer easily to this new problem. Although the direct impact of sentence trimming was minimal compared to other features used in the system, the interaction of the other features resulted in trimmed sentences accounting for nearly half of the selected summary sentences.

1 Introduction

This paper presents an initial application of UMD/BBN's single-document summarization approach (Trimmer), to the problem of multi-document summarization. Trimmer uses linguistically-motivated heuristics to trim syntactic constituents from sentences until a length threshold is reached. Trimmer was designed with the intention of compressing a lead sentence into a space consisting of tens of characters. Given that MSE2005 and DUC2005 required a longer summary based on inputs from different sources, we investigated the feasibility of applying sentence trimming approach to multi-document summarization.

We incorporated sentence trimming into a feature-based summarization system, called Multi-Document Trimmer (MDT), by using sentence trimming as both a pre-processing stage and a feature for sentence ranking. Trimmer is used to pre-process the input documents, creating multiple partially trimmed sentences for each original sentence. The number of trimming operations applied to the sentence is used as a feature in the sentence ranker.

We demonstrate that we were able to port Trimmer easily to this new problem. Although the impact of sentence trimming was minimal compared to other features used in the system, the interaction of the other features resulted in trimmed sentences accounting for nearly half of the selected summary sentences.

The next section relates our approach to other existing summarization systems. Following this, we describe the MDT approach and then present the results of running our system in the DUC2005 task.

2 Background

A successful approach to extractive multi-document summarization is to rank candidate sentences according to a set of factors, iteratively reranking to avoid redundancy within the summary. MEAD (Radev et al., 2004; Erkan and Radev, 2004) ranks documents according to a linear combination of features including centroid, position and first-sentence overlap. Once a set of sentences has been chosen as the summary, all sentences are rescored with a redundancy penalty based on word overlap with the chosen sentences. A new set of summary sentences is chosen based on the reranking. This is iterated until there are no changes in the summary. MDT differs in that syntactic trimming is used to provide shorter, but still grammatically correct, variants of the sentences as candidates. Also, MDT treats redundancy as a dynamic feature of unselected candidates.

Syntactic shortening has been used as in multi-document summarization in the SC system (Blair-Goldensohn et al., 2004). The SC system preprocesses the input to remove appositives and relative clauses. MDT differs from SC in that a wider variety of syntactic structures are candidates for trimming, and that multiple trimmed variants of each sentence are provided.

Minimization of redundancy is an important element of a multi-document summarization system. Carbonell and Goldstein (1998) propose Maximal Marginal Relevance (MMR) as a way of ranking documents found by an Information Retrieval system so that the front of the list will contain diversity as well as high relevance. Goldstein, Mittal, Carbonell and Kantrowitz (2000) demonstrate MMR applied to the problem multi-document summarization. MDT borrows the ranking approach of MMR, but uses a different set of features. MDT, like MEAD, uses feature weights that were optimized to maximize an automatic metric.

3 Multi-Document Trimmer

MDT consists of a three-stage process. First a syntactic trimmer is used to provide multiple trimmed versions of each sentence in each document of a topic set. Each of these trimmed variants is given a relevance score, either to a query if one is available, or to the topic set as a whole. Finally sentences are chosen according to a linear combination of features.

We used six features in ranking the candidate sentences.

- Fixed features
  - Position. The zero-based position of the sentence in the document.
  - Sentence Relevance. The relevance score of the sentence to the query or the topic set.
  - Document Relevance. The relevance score of the document to the topic set.
  - Trims. The number of trimmer rules applied to the sentence.
- Dynamic features
  - Redundancy. A measure of how similar the sentence is to the current state of the summary.
  - Sent-from-doc. The number of sentences already selected from the sentence's document.

The score for a sentence is a linear combination of these six features.

3.1 Syntactic Sentence Trimming

We use Trimmer (Dorr et al., 2003; Zajic et al., 2004) to provide multiple trimmed versions of the sentences in the documents. Trimmer uses linguistically-motivated heuristics to remove low-content syntactic constituents until a length threshold is reached. In the context of multi-document summarization, each intermediate stage of trimming is presented as a potential summary sentence.

The following example shows the behavior of Trimmer as trimming rules are applied sequentially to a sentence from the MSE2005 test set. The first example is the original sentence. In each example, the constituent to be removed next is shown in italics. Ideally, each application of a trimming rule yields a grammatical sentence.

(1) after 15 years and an investigation involving thousands of interviews, canada's police have arrested the men they say masterminded the deadliest-ever bombing *of an airplane*.

(2) after 15 years and an investigation involving thousands *of interviews*, canada's police have arrested the men they say masterminded the deadliest-ever bombing.

(3) *after 15 years and an investigation involving thousands,* canada's police have arrested the men they say masterminded the deadliest-ever bombing.

(4) canada's police have arrested the men *they say masterminded the deadliest-ever bombing*.

(5) canada's police have arrested the men.

MDT excludes certain document-initial material from the summary. In particular, datelines from written news and low-content introductory sentences from broadcast news. The Trimmer component of MDT identifies the first content sentence of a document as the first sentence containing six or more words. It does not generate trimmed or untrimmed versions of any sentences that precede the first content sentence.

The Trimmer component of MDT also differs from single document Trimmer in that punctuation is preserved from the original document. In the context of single document headline generation, punctuation was entirely removed from headlines. Punctuation took up character space, and the removal of punctuation usually did not interfere with human understanding of the generated headlines. In the context of multi-document summarization, the inclusion of punctuation does not take up space, because summary size is measured in words, not characters. Also, punctuation has a much larger effect on the readability of the summaries.

3.2 Sentence Relevance Scoring

The relevance score is broken down into two separate components: the matching score between a trimmed sentence and the query, and a similarity score between the document containing the trimmed sentence in question and the entire cluster of relevant documents. We assume that sentences having higher term overlap with the query and sentences originating from documents more "central" to the topic cluster are preferred for inclusion in the final summary.

The matching score between a trimmed sentence and the query is an *idf*-weighted count of overlapping terms (number of terms shared by the two text segments). Inverse document frequency (*idf*), a commonly-used measure in the information retrieval literature, can roughly capture the salience terms. The *idf* of a term *t* is defined by $log(N/c_t)$, where N is the total number of documents in a particular corpus, and c_t is the number of documents containing term *t*; these statistics were calculated from one year's worth of LA Times articles. Weighting term overlap by inverse document frequency captures the intuition that matching certain terms is more important than matching others.

The similarity between a particular document and the cluster of relevant documents was calculated using Lucene, a freely-available off-the-shelf information retrieval system. This basic intuition is that certain documents are more "central" to the topic at hand; all things being equal, sentences from such documents should be preferred. This similarity score is the average of the document's similarity with every relevant document in the cluster (as measured by Lucene's built-in comparison function). In order to obtain an accurate distribution of term frequencies to facilitate the similarity calculation, we indexed all relevant documents along with a comparable corpus (one year of the LA Times)—this additional text essentially serves as a background model for non-relevant documents.

3.3 Redundancy Scoring

To measure how redundant a sentence is with respect to the current state of the summary, we imagine that a candidate sentence has been generated from a combination of the current state of the summary and the general language. The parameter λ denotes the probability that a word from the candidate was generated by the current summary, and $(1 - \lambda)$ is the probability that the word was generated by the general language. We have set $\lambda = 0.3$ as a conventional starting value, but have not yet tuned this parameter.

Suppose that a candidate is fully redundant to the current summary. Then the probability that a word w occurs in the candidate is

$$P(w) = \lambda P(w|D) + (1 - \lambda)P(w|C)$$

where D is the current state of the summary and C is the corpus (in this case, the concatenation of all the documents in the topic set). We calculate the probabilities by counting the words in the current summary and the documents of the topic set:

$$P(w|D) = \frac{count\ of\ w\ in\ D}{size\ of\ D}$$

$$P(w|C) = \frac{count\ of\ w\ in\ C}{size\ of\ C}$$

We take the probability of a sentence to be the product of the probabilities of its words, so we calculate redundancy as:

$$Redundancy(S) = \prod_{s \in S} \lambda P(s|D) + (1-\lambda)P(s|C)$$

For ease of computation, we actually use log probabilities:

$$\sum_{s \in S} \log(\lambda P(s|D) + (1-\lambda)P(s|C))$$

If a candidate sentence is truly redundant to the current summary, it will have a relatively high probability of having been "generated" in this way. If it is non-redundant it will have a low probability.

Prior to calculating the redundancy score, we remove stopwords and apply the Porter Stemmer (Porter, 1980) to the sentence, the current summary and the corpus.

3.4 Sentence Selection

The score for a sentence is a linear combination of the six features described above. The highest ranking sentence from the pool of eligible candidates is chosen for inclusion in the summary. When a candidate is chosen, all other trimmed variants of that sentence are eliminated. After a sentence is chosen, the dynamic features, redundancy and sent-from-doc, are re-calculated, and the candidates are re-ranked. Sentences are added to the summary until the space is filled. Once the space is filled, the sentences of the summary are re-ordered so that sentences from the same document occur together, in the same relative order that they occurred in the original document. The final sentence of the summary will be truncated if it goes over the word limit.

The weights for the factors were determined by manually optimizing on a set of training data to maximize the ROUGE-2 recall score (Lin and Hovy, 2003), using ROUGE version 1.5.5. MDT can be configured to prevent any trimmed sentences from appearing in the summary by setting the trim weight to $-\infty$.

Feature	Submitted Weight	Revised Weight
Position	-1	-10
Relevance	20	28
Trims	-2	$-\infty$
Redundancy	-20	-20
Sent-from-doc	-0.5	-3

Table 1: MSE2005 Tuned Feature Weights

4 MSE2005 Evaluation

The Multilingual Summarization Evaluation (MSE) 2005 task was to generate 100-word summaries for 25 clusters of documents. Each cluster of documents consisted of news stories about a single event, some originally in English and others translated into English from Arabic. The MSE2005 task did not include queries or topic descriptions for the topic clusters. We used a single relevance calculation, the relevance of the sentence to the topic cluster. The feature weights were manually optimized to maximize ROUGE-2 recall for the MSE2005 training data. MDT was run on the MSE2005 test data using these weights. The optimized weights for the submitted system are shown in Table 1.

Table 2 shows the ROUGE scores and relative ranks of the version of MDT that was submitted to the MSE2005 evaluation. After the evaluation, we ran an improved version of MDT that removed datelines from consideration and made some improvements to the trimming component on the same data. Feature weights were re-optimized to maximize ROUGE-2 recall of the revised MDT system for the MSE2005 training data, and the revised weights are also shown in Table 1. We found that the optimal weight for trimming rules in the submitted system was -2, but the optimal weight for the revised system was $-\infty$, which would prevent any trimmed variants from appearing in the summary. We ran the revised MDT on the MSE2005 test data with three settings of trim weight: $-\infty$, -2 and 0. A weight of zero effec-

ROUGE	Avg Recall	Avg Precision	Avg F
1	0.39780 (10)	0.40926 (15)	0.40340 (12)
2	0.11849 (16)	0.12193 (15)	0.12017 (18)
3	0.04821 (20)	0.04971 (21)	0.04894 (20)
4	0.02500 (24)	0.02586 (24)	0.02542 (24)
L	0.35424 (11)	0.35506 (15)	0.35004 (12)
W-1.2	0.11740 (11)	0.22702 (15)	0.15472 (12)
SU4	0.14971 (16)	0.15420 (18)	0.15190 (18)

Table 2: ROUGE scores for MDT (System 19) in MSE2005, with ranks out of 27 automatic systems

Trim Weight	Avg R-1 Recall	Avg R-2 Recall	Avg R-SU4 Recall
0	0.39062	0.13193	0.15617
-2	0.40287	0.13143	0.16122
$-\infty$	0.40477	0.13039	0.16122

Table 3: Avg ROUGE recall scores for revised system on MSE2005 data

tively removes the number of trims as a ranking factor. The ROUGE-2 average recall for these runs are shown in Table 3.

The ROUGE scores for the revised MDT on the MSE2005 data show that the use of trimming increases the ROUGE-2 score by a small, non-significant amount, even though the optimized weight on trim rules for the training data was $-\infty$. However the ROUGE-1 and ROUGE-SU4 scores rank the three weights in different orders, again with non-significant differences. This suggests that 24 data points (the number of topics in the MSE2005 training data) may not be sufficient to optimize 5 independent factors, and that the ROUGE scores do not show whether the use of trimming improved the performance of MDT on the MSE2005 test data or not.

The MSE submissions were also evaluated using the Pyramid method (Nenkova and Passonneau, 2004). Peer and model summaries are manually searched for summarization content units (SCUs) of differing importance. The more important or central SCUs carry more weight in the scoring. The Pyramid score measures the proportion of good SCUs in the summary, and corresponds to precision. The modified pyramid score is the proportion of good SCUs to the number of good SCUs that one would expect in a summary of the given size. Modified pyramid scores correspond to recall. MDT got the third highest Pyramid score of ten peer systems on the MSE 2005 task, and was fifth out of ten for the modified pyramid score.

Feature	Weight	Avg for Summ. Sents	Avg for All Sents
Position	-3	0.5781	24.26
Sent Rel	0.05	11.20	4.533
Doc Rel	35	0.3508	0.07333
Trims	0	0.9983	1.806
Redundancy	-5		
Sent from Doc	-0.75		

Table 4: DUC2005 Feature Weights, and average values of static features for candidate sentences in the DUC2005 test data.

5 DUC2005 Evaluation

The DUC2005 task was to generate 250-word summaries for 50 sets of documents. The members of each document set were selected to contain information about a topic query, even though the documents might not be primarily about the topic. The summaries were to focus on information relevant to the topic query. The feature weights for the six features were manually optimized to maximize the ROUGE-2 recall score on the eleven DUC2005 sample topics, using reference summaries that we created. The feature weights are shown in Table 4, along with the average values of the features for the candidate sentences that were chosen to be in the summaries, and for the all of the candidate sentences from the DUC2005 test data.

The small positive weight on sentence relevance reflects the relatively large values of this factor among all the candidates. The small weight prevents sentence relevance from swamping the other factors. Similarly, the large positive weight on document relevance reflects the small values of that factor, and prevents the document relevance from being swamped by the other factors. Note that among all the candidates the average number of applied trimming rules is just under two, while for the selected sentences it is just under one rule. Even though the feature weight of zero means that the trim rule feature doesn't affect a sentence's score, the other factors combine to favor some trim rule applications and reject others, which is the desired behavior. The distribution of

N	Number of Summary Sentences with N Applied Trim Rules
0	314
1	137
2	43
3	29
4	24
5	13
6	11
7	4
10	1

Table 5: Distribution of number of trim rules applied to summary sentences in the DUC2005 test data

N	Number of Documents contributing N sents to a summary
0	1241
1	188
2	106
3	40
4	6
5	5
6	0
7	1

Table 6: Distribution of number of sentences contributed to a summary from a document for DUC2005 test data

ROUGE	Avg Recall	Avg Precision	Avg F
1	0.33940 (20)	0.32898 (31)	0.33403 (23)
2	0.05520 (24)	0.05360 (23)	0.05437 (25)
3	0.01298 (24)	0.01261 (29)	0.01279 (26)
4	0.00544 (25)	0.00528 (28)	0.00536 (25)
L	0.31331 (19)	0.30368 (29)	0.30835 (21)
W-1.2	0.09056 (19)	0.16232 (29)	0.11621 (20)
SU4	0.10970 (22)	0.10632 (31)	0.10796 (24)

Table 7: ROUGE scores for MDT (System 27), with ranks out of 32 automatic systems

Question	Avg Score	Rank
Grammaticality	2.83	40
Non-Redundancy	4.36	32
Referential Clarity	2.89	29
Focus	2.73	37
Structure & Coherence	1.57	42

Table 8: Linguistic scores for MDT (System 27) with ranks out of 42, including humans

number of trim rule applications in the selected sentences is shown in Table 5. Of the 576 candidates selected to appear in the DUC2005 test data summaries, 262 or 45% had at least one trimming rule applied to them.

The distribution of summary sentences from documents, shown in Table 6 shows that the negative Sentence from Document weight was effective at limiting the number of summary sentences selected from a single document.

Table 7 shows the ROUGE scores for MDT on the DUC2005 test data with ranks out of 32 submitted systems. MDT generally ranked higher for recall than for precision, suggesting that MDT is currently more successful at finding relevant content than it is at weeding out irrelevant content.

The DUC2005 evaluation also included human judgments of linguistic quality and responsiveness to the query. The scores and ranks for MDT on these human evaluations are shown in Tables 8 and 9. We believe that the extremely low score for grammaticality reflects the fact that trimmed sentences were actually getting into the summaries. Although Trimmer attempts to preserve grammaticality, it is to be expected that Trimmer will not preserve grammaticality as well as simply extracting sentences and leaving them alone. In the area of non-redundancy, MDT scored well, but so did most other systems, indicating that non-redundancy is not a difficult property to achieve. The low scores in coherence and referential clarity correctly reveal that MDT does not yet have any mechanism for dealing with units larger than the sentence.

A pyramid evaluation was done on 28 of the DUC2005 submissions. Table 10 shows the performance of MDT in the DUC2005 pyramid evaluation. MDT ranked higher among submitted systems in the MSE2005 Pyramid evaluation than in the DUC2005 Pyramid evaluation. The DUC2005 task differed from the MSE2005 task in two important ways: the summary length was longer (250

	Avg Score	Rank
Including humans	15.64	32 of 42
Not including humans	15.52	22 of 32

Table 9: Average Scaled Responsiveness scores for MDT (System 27) with ranks

	Avg Score	Rank
Pyramid Score	0.17453	22 of 28
Modified Pyramid Score	0.14015	22 of 28

Table 10: Average Pyramid scores for MDT (System 27) with ranks

vs 100 words) and the topic clusters were generated by a query. MDT may be doing a better job at identifying primary content than at distinguishing among possible secondary content. In this case, it would fill a shorter summary with good content, but include irrelevant content for much longer summaries. The bias in favor of high-relevance sentences appearing at the front of a document should be weaker if there is a query or topic description available, as in the DUC2005 task. However, the query-focused version of MDT used in DUC2005 was *less* likely to select a sentence far from the start of the document than the non-query focused version used for MSE2005. The average story position of a selected candidate was 0.55781 for DUC2005 and 1.211 for MSE2005, probably a result of the difference in the optimized weight for the position factor (-3 for DUC2005 and -1 for MSE2005).

6 Conclusion and Future Work

We believe that the MSE2005 and DUC2005 evaluations of Multi-Document Trimmer indicate that sentence compression by syntactic trimming can be a useful tool in the context of multi-document summarization. MDT must be augmented with modules to improve summary coherence and structure.

The current state of MDT represents our first effort to incorporate syntactic trimming into multi-document summarization. We plan to analyze MDT's errors to determine why it did not select trimmed versions of sentences that actually removed unimportant syntactic constituents. We also plan to examine the output of the syntactic trimmer to determine if it is actually providing appropriate alternatives to the original sentences.

Acknowledgments

The University of Maryland authors are supported, in part, by BBNT Contract 020124-7157, DARPA/ITO Contract N66001-97-C-8540, and NSF CISE Research Infrastructure Award EIA0130422.

References

Sasha Blair-Goldensohn, David Evans, Vasileios Hatzivassiloglou, Kathleen McKeown, Ani Nenkova, Rebecca Passonneau, Barry Schiffman, Andrew Schlaikjer, Advaith Siddharthan, and Sergey Siegelman. 2004. Columbia university at duc 2004. In *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pages 23–30.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop, Edmonton, Alberta, Canada*, pages 1–8.

Güneş Erkan and Dragomir R. Radev. 2004. The university of michigan at duc2004. In *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pages 120–127.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, , and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, pages 40–48.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrences Statistics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004, Boston*.

Martin Porter. 1980. An algorithm for suffix stripping. In *Program*, volume 14(3), pages 130–137.

Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. In *Information Processing and Management*, volume 40, pages 919–938.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2004. Bbn/umd at duc2004: Topiary. In *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pages 112–119.