In Proceedings of the 2006 Document Understanding Conference
(DUC-2006) at NLT/NAACL 2006, June 2006, New York, New York

Sentence Compression as a Component of a Multi-Document Summarization System

David M. Zajic
Department of Computer Science
University of Maryland
College Park, MD 20742 dmzajic@umiacs.umd.edu

Jimmy Lin
University of Maryland
jimmylin@umiacs.umd.edu

Bonnie J. Dorr
University of Maryland
bonnie@cs.umd.edu

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

Abstract

We applied a single-document sentence-trimming approach (Trimmer) to the problem of multi-document summarization. Trimmer was designed with the intention of compressing a lead sentence into a space consisting of tens of characters. In our Multi-Document Trimmer (MDT), we use Trimmer to generate multiple trimmed candidates for each sentence. Sentence selection is used to determine which trimmed candidates provide the best combination of topic coverage and brevity. We demonstrate that we were able to port Trimmer easily to this new problem. We also show that 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. Finally, we present an error analysis that shows that, while sentence compressions is making space for additional sentences, more work is needed in the area of generating and selecting the right candidates.

1 Introduction

This paper presents an application of UMD/BBN's single-document sentence-trimming 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. Multi-document summarization imposes a global length constraint rather than a sentence-level length constraint. We use Trimmer to generate multiple trimmed candidates for each sentence. Sentence selection is then used to determine which trimmed candidates provide the best combination of topic coverage and brevity.

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. We also show that 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. Finally, we present an error analysis that shows that, while sentence compressions is making space for additional sentences, more work is needed in the area of generating and selecting the right candidates.

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 DUC2006 task.

2 Background

A successful approach to extractive multi-document summarization is to rank candidate sentences according to a set of factors, iteratively re-ranking 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 re-ranking. 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 pre-processes the input to remove appositives and relative clauses. CLASSY (Conroy et al., 2005) uses an HMM sentence selection approach combined with a conservative sentence compression method based on shallow parsing to detect lexical cues to trigger phrase eliminations. MDT differs from SC and CLASSY in that a wider variety of syntactic structures are candidates for trimming, and that multiple trimmed candidates 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 candidates 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 eight 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.
  - Document Relevance. The relevance score of the document to the query.
  - Sentence Centrality. The centrality score of the sentence to the document.
  - Document Centrality. The centrality score of the document to the topic.
  - 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 from a parse tree until a length threshold for the surface string is reached. The DUC2006 submission used Charniak's parser (Charniak, 2000). 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. (Zajic et al., 2005) 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 multi-sentence summaries.

3.2 Sentence Relevance Scoring

The relevance score is broken down into four separate components: the matching score between a trimmed sentence and the query, the matching score between the document and the query, a similarity (or *centrality*) score between a sentence and the document in which it appears, 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 or document 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 sentence and document, or between a 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 candidates 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 $-\infty$.

4 DUC2006 Evaluation and Analysis

The DUC2006 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 DUC2005 test data, using the DUC2005 reference summaries. The feature weights are shown in Table 1.

In the DUC2006 evaluations, the UMD/BBN system was System 32. Table 2 shows the ROUGE scores for MDT on the DUC2006 test data with ranks out of 35 submitted systems. MDT gen-

Feature	Weight
Position	-3
Sentence Relevance	1.0
Document Relevance	1.0
Sentence Centrality	1.0
Document Centrality	1.0
Trim Operation Count	1.0
Redundancy	-2.0
Sentences from Document	-0.75

Table 1: DUC2006 Feature Weights

ROUGE	Avg Recall	Avg Precision	Avg F
1	0.38196 (17)	0.37617 (27)	0.37898 (20)
2	0.08051 (13)	0.07022 (32)	0.07985 (15)
3	0.02484 (9)	0.02493 (11)	0.02461 (10)
4	0.01100 (7)	0.01078 (9)	0.01088 (8)
L	0.35280 (16)	0.34748 (25)	0.35006 (19)
W-1.2	0.10316 (14)	0.18533 (23)	0.13250 (17)
SU4	0.13600 (13)	0.13388 (32)	0.13490 (16)

Table 2: ROUGE scores for MDT (System 32), with ranks out of 35 automatic systems

erally 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 DUC2006 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 3 and 4. 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 alonel 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.

Question	Avg Score	Rank
Grammaticality	2.74	44
Non-Redundancy	3.76	45
Referential Clarity	2.84	36
Focus	3.42	37
Structure & Coherence	1.84	32

Table 3: Linguistic scores for MDT (System 32) with ranks out of 45, including humans

	Avg Score	Rank
Content	2.6	13
Overall	2.08	23

Table 4: Average Responsiveness scores for MDT (System 32) with ranks out of 35 automatic systems

The intuition behind the use of sentence compression in multi-document summarization is that by removing non-relevant constituents from summary sentences, we can make room for additional relevant sentences within the length constraint. Consider three phenomena we would expect to see when a multi-document summarization system is augmented with sentence compression.

- A net increase in the average number of sentences per summary.
- The shortening of some summary sentences by the removal of non-relevant constituents.
- The addition of some relevant sentences.

We ran the summary generator with the constraint that it could select only original source sentences, to serve as a basis of comparison with the submitted DUC2006 system. The average summary without sentence compression contained 7.6 sentences. The average summary with sentence compression contained 11.6 sentence. On average, the use of sentence compression caused 1.94 sentences to be dropped from each summary, and 5.90 new sentences to be added, for a net gain of 3.96 sentences. Of the orignal sentences in the summaries without compression, 21.5% appeared unchanged in the summaries with compression, and 53.1% appeared with some constituents removed.

Consider Topic D0602, which concerns the use of steroids by female athletes. The first sentences of the summaries show the operation of Trimmer in the context of multi-document summarization.

(6) Without Trimmer: Two Moroccan female athletes have been stripped of gold and bronze medals for using a muscle-building steroid in the first reported cases of doping at the Arab Games, an official said Friday.

(7) With Trimmer: Two Moroccan female athletes have been stripped of gold and bronze medals for using a muscle-building steroid.

However, sometimes Trimmer results in an anomalous sentence.

(8) Without Trimmer: Medical experts say athletes frequently take anabolic steroids in doses high enough to have dangerous consequences, with some users known to have taken 10 to 100 times the recommended dosage.

(9) With Trimmer: Athletes frequently take anabolic steroids in doses high enough to have dangerous consequences, with some users.

In this case the original sentence was not fully relevant to the query, and with trimming it is still not fully relevant but it takes up less space.

The sentences that are selected to fill the available space are sometimes fully relevant, but sometimes not. Consider two of the trimmed sentences that were added for Topic D0602.

(10) 2.4 percent of female high-school students acknowledged using the illegal steroids.

(11) Its potential for inducing serious side effects is similar to that of anabolic steroids.

Of the seven sentences added by the use of sentence compression only one was fully relevant to the query. Moreover, the one sentence which was dropped entirely from summary that didn't use compression was fully relevant:

(12) Dr. Charles E. Yesalis, a steroid authority at Pennsylvania State University, was among the first to analyze the 1997 female-adolescent data in December of that year in the Archives of Pediatrics and Adolescent Medicine and to sound the alarm.

It appears that Trimmer is operating as intended in the context of multi-document summarization, but that the expected benefit is not being fully realized because the summary generator is not yet able to make good use of the extra available space.

5 Conclusion and Future Work

Our system uses sentence compression in the context of multi-document summarization to generate multiple trimmed candidates for each sentence in the source texts. In order for this approach to succeed three things must happen. First, the sentence compression system must generate among its output some trimmed candidates which are grammatical, coherent and which preserve the right information. Second, the candidate choosers must be able to select these right candidates over the original sentences or degenerate alternative trims. Finally, the candidate choosers must make good use of the extra space to choose query-relevant and non-redundant additional candidates.

The error analysis has shown that while sentence compression is making space for additional sentences, more work is needed in the area of generating and selecting the right candidates.

Acknowledgments

This work has been supported, in part, under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001, the TIDES program of the Defense Advanced Research Projects Agency, BBNT Contract No. 9500006806, and the University of Maryland Joint Institute for Knowledge Discovery. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

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.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the North American ACL(NAACL)*, pages 132–139.

John M. Conroy, Judith D. Schlesinger, and Jade Goldstein Stewart. 2005. Classy query-based multi-document summarization. In *Proceedings of the 2005 Document Understanding Workshop, Boston*.

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.

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.

David M. Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2005. UMD/BBN at MSE2005. In *Proceedings of the MSE2005 track of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Meatures for MT and/or Summarization, Ann Arbor, MI*.