BBN/UMD at DUC-2004: Topiary

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

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

Abstract

This paper reports our results at DUC-2004 and describes our approach, implemented in a system called Topiary. We will show that the combination of linguistically motivated sentence compression with statistically selected topic terms performs better than either alone, according to some automatic summary evaluation measures.

1 Introduction

This paper reports our results at DUC-2004 and describes our approach, implemented in a system called Topiary. Our algorithm combines sentence compression with Unsupervised Topic Discovery (UTD). Our sentence compression algorithm involves removing constituents from a parse tree of the lead sentence according to a set of linguistically-motivated heuristics until a length threshold is reached. UTD is a statistical method for deriving a set of topic models from a document corpus, assigning meaningful names to the topic models, and associating sets of topics with specific documents. The topics and sentence compressions are combined in a manner that preserves the advantages of each approach: the fluency and event-oriented information from the lead sentence with the broader coverage of the topic models.

The next section presents previous work in the area of automatic summarization. Following this we describe Hedge Trimmer and UTD in more detail, and describe the algorithm for combining sentence compression with topics. Next we show that Topiary scores higher than either Hedge Trimmer or UTD alone according to certain automatic evaluation tools for summarization. Finally we discuss the performance of Topiary in the DUC-2004 evaluations.

2 Previous Work

Our sentence compression algorithm is based on linguistically-motivated heuristics. Previous work on sentence compression (Knight and Marcu, 2000) uses a noisy-channel model to find the most probable short string that generated the observed full sentence. Other work (Euler, 2002) combines a word-list with syntactic information to decide which words and phrases to cancel. Our approach differs from Knight's in that we do not use a statistical model, so we do not require any prior training on a large corpus of story/headline pairs. Topiary shares with Euler the combination of topic lists and sentence compression. However Euler uses the topic lists to guide sentence selection and compression towards a query-specific summary, whereas Topiary uses topics to augment the concept coverage of a generic summary.

Summaries can also consist of lists of words or short phrases indicating that the topic or concept they denote is important in the document. Extractive topic summaries consist of keywords or key phrases that occur in the document. (Bergler et al., 2003) achieves this by choosing noun phrases that represent the most important text entities, as represented by noun phrase coreference chains. (Zhou and Hovy, 2003) imposes fluency onto a topic list by finding phrase clusters early in the text that contain important topic words found throughout the text. In text categorization documents are assigned to pre-defined categories. This is equivalent to assigning topics to a document from a static topic list, so the words in the summary need not actually appear in the document. (Lewis, 1992) describes a probabilistic feature-based method for assigning Reuters topics to news stories. OnTopic (Schwartz et al., 1997) uses a HMM to assign topics from a topic-annotated corpus to a new document.

3 Algorithm Description

Topiary produces headlines by modifying the output of Hedge Trimmer, a sentence compression algorithm, to leave enough space for some topic words and phrases, provided by UTD. In this section we will give brief descriptions of Hedge Trimmer, recent modifications to Hedge Trimmer, and UTD. We will then describe how Hedge Trimmer and UTD are combined.

3.1 Hedge Trimmer

Hedge Trimmer (Dorr et al., 2003b) generates a headline for a news story by compressing the *lead* (or main) topic sentence according to a linguistically motivated algorithm. For text news stories, the first sentence of the document is taken to be the lead sentence. The compression begins by using the BBN SIFT parser (Miller et al., 1998) to parse the lead sentence, and BBN IdentiFinder™(Bikel et al., 1999) to detect named entities and time expressions. Then low-content syntactic constituents are removed. Some constituents, such as certain determiners (the, a) and time expressions are always removed, because they rarely occur in human-generated headlines and are low-content in comparison to other constituents. Other constituents are removed one-by-one until a length threshold has been reached. These include, among others, relative clauses, verb-phrase conjunction, preposed adjuncts and prepositional phrases that do not contain named entities. [1] The threshold can be specified either in number of words or number of characters. If the threshold is specified in number of characters, Hedge Trimmer will not include partial words.

[1] More details of the Hedge Trimmer algorithm can be found in (Dorr et al., 2003b) and (Dorr et al., 2003a).

3.2 Recent Hedge Trimmer Work

Recently we have investigated a rendering of the summary as "Headlinese" (Mårdh, 1980) in which certain constituents are dropped with no loss of meaning. The result of this investigation has been used to enhance Hedge Trimmer, most notably the removal of certain instances of *have* and *be*. For example, the previous headline generator produced summaries such as Sentence (2), whereas the have/be removal produces Sentence (3).

(1) <u>Input:</u> At least 231 people have been confirmed dead in Honduras from former-hurricane Mitch, bringing the storm's death toll in the region to 357, the National Emergency Commission said Saturday.

(2) <u>Without participle have/be removal:</u> At least 231 people have been confirmed dead bringing storm's death toll

(3) <u>With participle have/be removal:</u> At least 231 people confirmed dead in Honduras bringing storm's death toll

Have and *be* are removed if they are part of a past or present participle construction. In this example, the removal of *have been* allows a high-content constituent *in Honduras* to fit into the headline.

The removal of forms of *to be* allows Hedge Trimmer to produce headlines that concentrate more information in the allowed space. The removal of forms of *to be* results in sentences that are not grammatical in general English, but are typical of Headlinese English. For example, Sentences (5), (6) and all other examples in this paper were trimmed to fit in 75 characters.

(4) <u>Input:</u> Russian space experts were making final preparations Thursday at the Baikonur rocket base to launch the first component of a multibillion dollar international space station after a year of delay.

(5) <u>Without *to be* removal:</u> Russian space experts were making final preparations

(6) With *to be* removal: Russian space experts making final preparations at Baikonur rocket base

When *have* and *be* occur with a modal verb, the modal verb is also removed. Sentence (9) shows an example of this. It could be argued that by removing modals such as *should* and *would* the meaning is vitally changed. The intended use of the headline must be considered. If the headlines are to be used for determining query relevance, removal of modals may not hinder the user while making room for additional high-content words may help.

(7) Input: Famine-threatened North Korea's harvest will be no better this year than last and could be worse, a senior U.N. aid official said Saturday.

(8) Without Modal-Have/Be Removal: Famine threatened North Korea's harvest will be no better this year

(9) With Modal-Have/Be Removal: Famine threatened North Korea's harvest no better this year than last

In addition when *it* or *there* appears as a subject with a form of *be* or *have*, as in extraposition (*It was clear that the thief was hungry*) or existential clauses (*There have been a spate of dog maulings*), the subject and the verb are removed.

Finally, for situations in which the length threshold is a hard constraint, we added some emergency shortening methods which are only to be used when the alternative is truncating the headline after the threshold, possibly cutting the middle of a word. These include removal of adverbs and adverbial phrases, adjectives and adjective phrases, and nouns that modify other nouns.

The main benefit of have/be removal is that it often shortens a headline by five to eight characters, without losing any content and rarely causing the sentence to become ungrammatical as Headlinese. Sometimes this shortening is enough to allow another constituent or, as we discuss in Section 3.4, an additional topic word or phrase to fit under the length threshold.

3.3 Unsupervised Topic Discovery

Unsupervised Topic Discovery (UTD) is used when we do not have a corpus annotated with topics. It takes as input a large unannotated corpus in any language and automatically creates a set of topic models with meaningful names. The algorithm has several stages. First, it analyzes the corpus to find strings of words that occur frequently. (It does this using a Minimum Description Length criterion.) These are frequently phrases that are meaningful names of topics.

Second, it finds the high-content phrases in each document (using a modified tf.idf measure). These are possible topic names for each document. It keeps only those names that occur in at least four different documents. These are taken to be an initial set of topic names.

In the third stage UTD trains topic models corresponding to these topic names. The modified EM procedure of OnTopic™is used to determine which words in the documents often signify these topic names. This produces topic models.

Fourth, these topic models are used to find the most likely topics for each document. This often adds new topics to documents, even though the topic name did not appear in the document. It also frequently removes topics that are not supported by the rest of the story.

We found, in various experiments, that the topics derived by this procedure were usually meaningful and that the topic assignment was about as good as when the topics were derived from a corpus that was annotated by people. We have also used this procedure on different languages and shown the same behavior.

Sentence (10) is a topic list generated for a story about the investigation into the bombing of the U.S. Embassy in Nairobi on August 7, 1998.

(10) BIN_LADEN EMBASSY BOMBING POLICE OFFICIALS PRISON HOUSE FIRE KABILA

3.4 Combination of Hedge Trimmer and Topics: Topiary

The Hedge Trimmer algorithm is constrained to take its headline from a single sentence. It is often the case that there is no single sentence that contains all the important information in a story. The information can be spread over two or three sentences, with pronouns or ellipsis used to link them. In addition, our algorithms do not always select the ideal sentence and trim it perfectly.

Topics alone also have drawbacks. UTD rarely generates any topic names that are verbs. Thus topic lists are good at indicating the general subject are but rarely give any direct indication of what events took place.

Topiary is a modification of the enhanced Hedge Trimmer algorithm to take a list of topics with relevance scores as additional input. The compression threshold is lowered so that there will be room for the highest scoring topic term that isn't already in the headline. This amount of threshold lowering is dynamic, because the trimming of the sentence can remove a previously ineligible high-scoring topic term from the headline. After trimming is complete, additional topic terms that do not occur in the headline are added to use up any remaining space.

This often results in one or more main topics about the story and a short sentence that says what happened concerning them. The combination is often more concise than a fully fluent sentence and compensates for the fact that the topic and the description of what happened to it do not appear in the same sentence in the original story.

Sentences (11) and (12) are the output of Hedge Trimmer and Topiary for the same story for which the topics in Sentence (10) were generated.

(11) FBI agents this week began questioning relatives of the victims

(12) BIN_LADEN EMBASSY BOMBING: FBI agents this week began questioning relatives

4 Evaluation

We used two automatic evaluation systems, BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003), to evaluate nine variants of our headline generation systems. Both measures make n-gram comparisons of the candidate systems to a set of reference summaries. In our evaluations four reference summaries for each document were used. The reference summaries were provided to the DUC participants by NIST. The nine variants were run on 489 stories from the DUC2004 single-document summarization headline generation task. The threshold was 75 characters, and longer headlines were truncated to 75 characters. We also evaluated a baseline that consisted of the first 75 characters of the document. The systems and the average lengths of the headlines they produced are shown in Table 1. Trimmer headlines tend to be shorter than the threshold because Trimmer removes constituents until the length is below the threshold. Sometimes it must remove a large constituent in order to get below the threshold. Topiary is able to make full use of the space by filling in topic words.

System	Description	Words	Chars
Trim	Trimmer no have/be removal no emergency shortening	8.7	57.3
Trim.E	Trimmer no have/be removal emergency shortening	8.7	57.1
Trim.HB	Trimmer have/be removal no emergency shortening	8.6	57.7
Trim.HB.E	Trimmer have/be removal emergency shortening	8.6	57.4
Top	Topiary no have/be removal no emergency shortening	10.8	73.3
Top.E	Topiary no have/be removal emergency shortening	10.8	73.2
Top.HB	Topiary have/be removal no emergency shortening	10.7	73.2
Top.HB.E	Topiary have/be removal emergency shortening	10.7	73.2
UTD	UTD Topics	9.5	71.1

Table 1: Systems and Headline Lengths

4.1 ROUGE

ROUGE is a recall-based measure for summarizations. This automatic metric counts the number of n-grams in the reference summaries that occur in the candidate and divides by the number of n-grams in the reference summaries. The size of the n-grams used by ROUGE is configurable. ROUGE-*n* uses 1-grams through *n*-grams. ROUGE-L is based on longest common subsequences, and ROUGE-W-1.2 is based on weighted longest common subsequences with a weighting of 1.2 on consecutive matches of length greater than 1.

The ROUGE scores for the nine systems and the baseline are shown in Table 2. Under ROUGE-1 the Topiary variants scored significantly higher than the Trimmer variants, and both scored significantly higher than the UTD topic lists with 95% confidence. Since fluency is not measured at all by unigrams, we must conclude that the Trimmer headlines, by selecting the lead sentence, included more or better topic words than UTD. The highest scoring UTD topics tend to be very meaningful while the fifth and lower scoring topics tend to be very noisy. Thus the higher scores of Topiary can be attributed to including only the best of the UTD topics while preserving the lead sentence topics. The same groupings occur with ROUGE-L and ROUGE-W, indicating that the longest common subsequences are dominated by sequences of length one.

Under the higher order ROUGE evaluations the systems group by the presence or absence of have/be removal, with higher scores going to systems in which have/be removal was performed. This indicates that the removal of these light content verbs makes the summaries more like the language of headlines. The value of emergency shortening over truncation is not clear.

4.2 BLEU

BLEU is a system for automatic evaluation of machine translation that uses a modified n-gram precision measure to compare machine translations to reference human translations. This automatic metric counts the number of n-grams in the candidate that occur in any of the reference summaries and divides by the number of n-grams in the candidate. The size of the n-grams used by BLEU is configurable. BLEU-*n* uses 1-grams through *n*-grams. In our evaluation of headline generation systems, we treat summarization as a type of translation from a verbose language to a concise one, and compare automatically generated headlines to human generated headlines.

The BLEU scores for the nine systems and the baseline are shown in Table 3. For BLEU-1

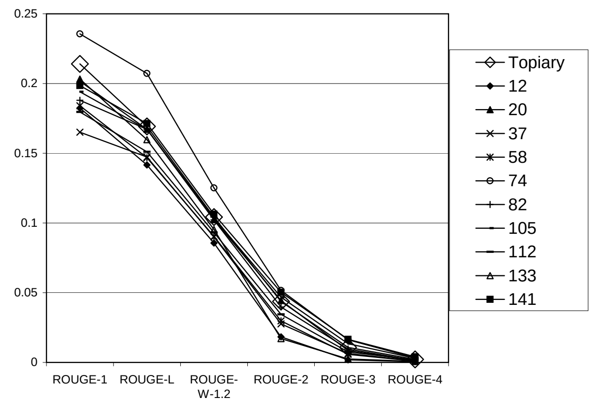

Figure 2: ROUGE Scores for DUC2004, Task 3, Machine Translations

the Topiary variants score significantly better than the Trimmer variants with 95% confidence. Under BLEU-2 the Topiary scores are higher than the Trimmer scores, but not significantly. Under BLEU-4 the Trimmer variants score slightly but not significantly higher than the Topiary variants, and at BLEU-3 there is no clear pattern. Trimmer and Topiary variants score significantly higher than UTD for all settings of BLEU with 95% confidence.

4.3 Performance in DUC-2004

We submitted Topiary output to the 2004 Document Understanding Conference Workshop for tasks 1 and 3. Figure 1 shows how Topiary peformed in comparison with other DUC2004 participants on task 1, using ROUGE. Task 1 was to produce a summary for a single news document no more than than 75 characters. The different ROUGE variants are sorted by overall performance of the systems. There was a wide range of performance among the submitted systems. Topiary scored highest among the automatic systems for the ROUGE-1, -2, -L and -W-1.2 measures, and scored second highest for the ROUGE-3 and -4 measures.

Task 3 was to produce a summary in English no longer than 75 characters for an Arabic document. Topiary output was submitted for two of the subtasks for task 3. In the first subtask, Topiary was given machine translations into English of the Arabic documents as input. In the second subtask Topiary was given manual translations of the Ara-

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-W-1.2
Top.HB.E	0.24914	0.06449	0.02122	0.00712	0.19951	0.11891
Top.HB	0.24873	0.06595	0.02267	0.00826	0.20061	0.11970
Top.E	0.24812	0.06169	0.01874	0.00562	0.19856	0.11837
Top	0.24621	0.06309	0.01995	0.00639	0.19856	0.11861
baseline	0.22136	0.06370	0.02118	0.00707	0.11738	0.16955
Trim.HB.E	0.20415	0.06571	0.02527	0.00950	0.18506	0.11127
Trim.HB	0.20380	0.06565	0.02508	0.00945	0.18472	0.11118
Trim.E	0.20105	0.06226	0.02221	0.00774	0.18287	0.11003
Trim	0.20061	0.06283	0.02266	0.00792	0.18248	0.10996
UTD	0.15913	0.01585	0.00087	0.00000	0.13041	0.07797

Table 2: ROUGE Scores sorted by ROUGE-1

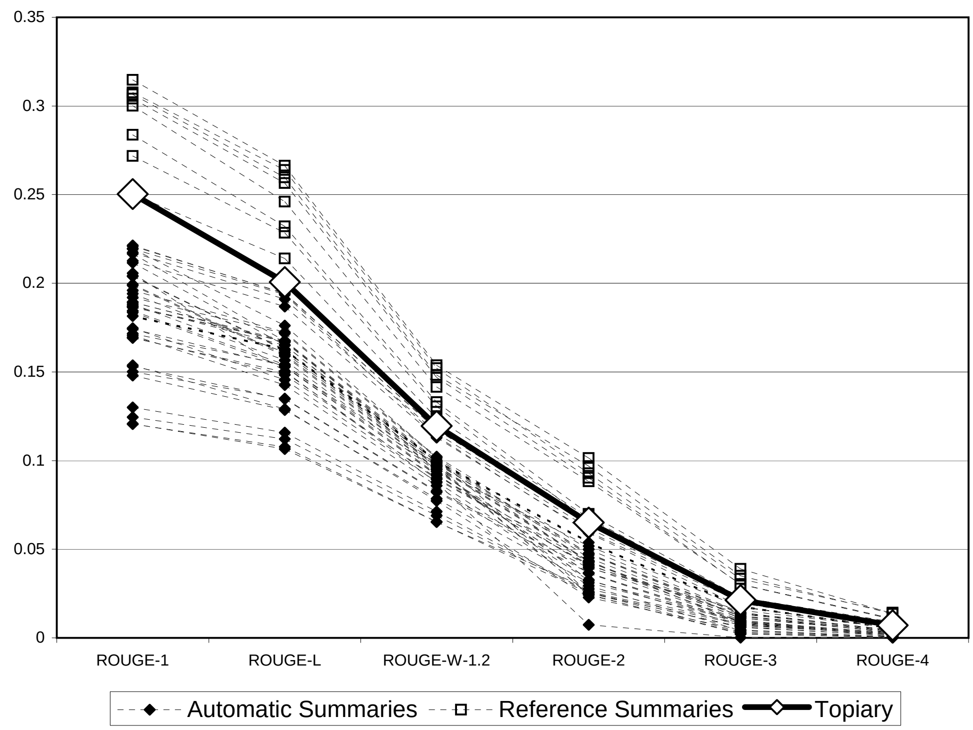

Figure 1: ROUGE Scores for DUC2004 Automatic Summaries, Reference Summaries and Topiary

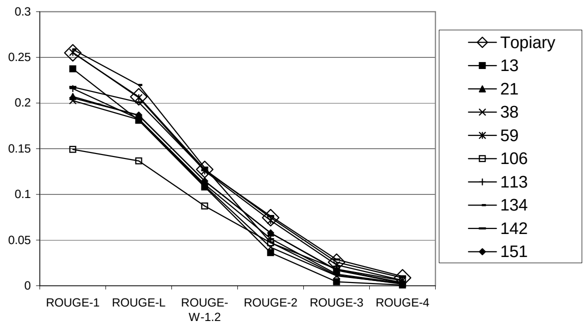

Figure 3: ROUGE Scores for DUC2004, Task 3, Manual Translations

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Top.HB.E	0.4368	0.2443	0.1443	0.0849
Top.HB	0.4362	0.2463	0.1476	0.0885
Top.E	0.4310	0.2389	0.1381	0.0739
Top	0.4288	0.2415	0.1417	0.0832
Trim.HB.E	0.3712	0.2333	0.1495	0.0939
Trim.HB	0.3705	0.2331	0.1493	0.0943
baseline	0.3695	0.2214	0.1372	0.0853
Trim.E	0.3636	0.2285	0.1442	0.0897
Trim	0.3635	0.2297	0.1461	0.0922
UTD	0.2859	0.0954	0.0263	0.0000

Table 3: BLEU Scores sorted by BLEU-1

bic documents. The ROUGE scores of the automatic systems for task 3 are shown in Figure 2 and Figure 3. The ROUGE scores of all systems increased with the improved fluency and topic coverage of the manual translations. It is meaningful to compare these sets of ROUGE scores because the same reference summaries were used for both subtasks. Topiary was among the high scoring systems for both subtasks.

5 Conclusions and Future Work

We have shown the effectiveness of combining sentence compression and topic lists to construct informative summaries. We have compared three approaches to automatic headline generation (Topiary, Hedge Trimmer and UTD) using two automatic summarization evaluation tools (BLEU and ROUGE). Topiary output was submitted to DUC-2004 for tasks 1 and 3, and was among the highest scoring systems for both tasks on all ROUGE measures.

We plan to perform a human study in which Topiary, Hedge Trimmer, UTD and other summarization methods will be evaluated on how well they help the subjects perform an extrinsic task. The extrinsic task will event tracking, in which subjects are asked to determine if a document is about a specific event. We also plan to extend the tools described in this paper to the domains of transcribed broadcast news and cross-language headline generation.

Acknowledgements

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

Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. 2003. Using knowledge-poor coreference resolution for text summarization. In *Proceedings of the 2003 Document Understanding Conference, Draft Papers*, pages 85–92, Edmonton, Candada.

D. Bikel, R. Schwartz, and R. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1/3).

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003a. Cross-language headline generation for hindi. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2:2.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003b. 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.

T. Euler. 2002. Tailoring text using topic words: Selection and compression. In *Proceedings of 13th International Workshop on Database and Expert Systems Applications (DEXA 2002), 2-6 September 2002, Aix-en-Provence, France*, pages 215–222. IEEE Computer Society.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization – step one: Sentence compression. In *The 17th National Conference of the American Association for Artificial Intelligence AAAI2000*, Austin, Texas.

David Lewis. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th annual international*

ACM SIGIR conference on Research and development in information retrieval, pages 37–50, Copenhagen, Denmark.

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.

Ingrid Mårdh. 1980. *Headlinese: On the Grammar of English Front Page Headlines*. Malmo.

S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, and R. Weischedel. 1998. Algorithms that Learn to Extract Information; BBN: Description of the SIFT System as Used for MUC-7. In *Proceedings of the MUC-7*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of Association of Computational Linguistics*, Philadelphia, PA.

R. Schwartz, T. Imai, F. Kubala, L. Nguyen, and J. Makhoul. 1997. A maximum likelihood model for topic classification of broadcast news. In *Eurospeech-97*, Rhodes, Greece.

Liang Zhou and Eduard Hovy. 2003. Headline summarization at isi. In *Proceedings of the 2003 Document Understanding Conference, Draft Papers*, pages 174–178, Edmonton, Candada.