

Comparing Topiary-Style Approaches to Headline Generation

Ruichao Wang¹, Nicola Stokes¹, William P. Doran¹, Eamonn Newman¹,
Joe Carthy¹, John Dunnion¹.

¹ Intelligent Information Retrieval Group, Department of Computer Science, University
College Dublin, Ireland.
{rachel, nicola.stokes, william.doran, eamonn.newman,
joe.carthy, john.dunnion}@ucd.ie

Abstract. In this paper we compare a number of Topiary-style headline generation systems. The Topiary system, developed at the University of Maryland with BBN, was the top performing headline generation system at DUC 2004. Topiary-style headlines consist of a number of general topic labels followed by a compressed version of the lead sentence of a news story. The Topiary system uses a statistical learning approach to finding topic labels for headlines, while our approach, the LexTrim system, identifies key summary words by analysing the lexical cohesive structure of a text. The performance of these systems is evaluated using the ROUGE evaluation suite on the DUC 2004 news stories collection. The results of these experiments show that a baseline system that identifies topic descriptors for headlines using term frequency counts outperforms the LexTrim and Topiary systems. A manual evaluation of the headlines also confirms this result.

1 Introduction

A headline is a very short summary (usually less than 10 words) describing the essential message of a piece of text. Like other types of summaries, news story headlines are used to help a reader to quickly identify information that is of interest to them in a presentation format such as a newspaper or a website. Although newspaper articles are always accompanied by headlines, there are other types of news text sources, such as transcripts of radio and television broadcasts, where this type of summary information is missing. In 2003 the Document Understanding Conference (DUC) [1] added the headline generation task to their annual summarisation evaluation. This task was also included in the 2004 evaluation plan, where summary quality was automatically judged using a set of n-gram word overlap metrics called ROUGE [2]. The best performing system at this workshop was the Topiary approach [3] which generated headlines by combining a set of topic descriptors extracted from the DUC 2004 corpus with a compressed version of the lead sentence of the news story, e.g. **COCHETEL CHECHNYA: French United Nations official kidnapped**

As can be seen these topic descriptors provide the reader with a general event description while the lead compressed sentence provides a more focussed summary of the news story.

Topiary-style summaries performed well in the ROUGE-based 2004 evaluation for a number of reasons. Firstly, summarisation researchers have observed that the lead sentence of a news story is in itself often an adequate summary of the text. However, it has also been observed that additional important information about a topic may be spread across other sentences in the news story. The success of the Topiary-style summaries at DUC 2004 can be attributed to fact that this technique takes both of these observations into consideration when generating titles.

In this paper, we compare three distinct methods of identifying topic labels and observe their effect on summary quality when combined with a compressed lead sentence. The Topiary system generates topic descriptors using a statistical approach called Unsupervised Topic Discovery (UTD) [3]. This technique creates topic models with corresponding topic descriptors for different news story events in the DUC 2004 corpus. One of the problems with this approach is that it requires additional on-topic documents related to the news story being summarised to facilitate the generation of relevant topic models and descriptors, i.e. Topiary used the DUC 2004 corpus when generating summaries for the DUC 2004 evaluation.

In this paper, we investigate the use of lexical cohesion analysis as a means of determining these event labels. The advantage of this approach is that the descriptors are gleaned from the source text being summarised, so no auxiliary training corpus or additional on-topic news story documents from the DUC corpus are needed to determine appropriate topic labels for a particular story headline. In Section 3 and 4, we describe the Topiary, and LexTrim (lexical cohesion-based approach) in more detail. The performance of these systems is compared with a baseline system called TFTrim (term frequency-based approach). These systems were evaluated using the ROUGE evaluation metrics on the DUC 2004 collection, and a manual evaluation performed by four human evaluators. The results of these experiments and our overall conclusions are discussed in Section 5 and 6 respectively. In the following section, we provide an overview of recent approaches to automatic headline generation described in the summarisation literature.

2 Related Work

The aim of this paper is to improve Zajic, Dorr and Schwartz's Topiary-style parse-and-trim approach to headline summarisation [3]. This approach falls into the extractive category of headline generation techniques, where a compressed sentence or series of compressed sentences are concatenated to produce a readable headline. Extractive approaches differ mainly in how they determine which textual units to include in the summary. Some common sentence weighting approaches include Kraaij et al.'s [4] probabilistic approach, Alfonseca et al.'s [5] genetic algorithmic approach, and Copeck et al.'s [6] approach based on the occurrence of features that denote appropriate summary sentences. These lexical, syntactic and semantic features

include the occurrence of discourse cues, the position of the sentence in the text, and the occurrence of content phrases and proper nouns. Biasing the extraction process with additional textual information such as these features is a standard approach to headline generation that has proved to be highly effective in most cases [4-7, 27].

At the DUC 2004 evaluation, a number of other parse-and-trim style headline techniques were presented [8-11]. However, all of these techniques were outperformed by the Topiary title generation system. More recently Zhou and Hovy [12] have proposed a template-based title generation approach, where part-of-speech tagged templates (automatically determined from a training corpus) are filled with content words selected using a keyword clustering technique. These templates help preserve the readability of the headlines by guiding the most suitable combination of keywords using grammatical constraints.

An alternative to extractive gisting approaches is to view the title generation process as being analogous to statistical machine translation. Wittbrock and Mittal's paper on 'ultra-summarisation' [13], was one of the first attempts to generate headlines based on statistical learning methods that make use of large amounts of training data. More specifically, during title generation a news story is 'translated' into a more concise version using the Noisy Channel model. The Viterbi algorithm is then used to search for the most likely sequence of tokens in the text that would make a readable and informative headline. This is the approach adopted by Banko et al. [14], Jin and Hauptmann [15], Berger and Mittal [16], and Zajic and Dorr's DUC 2002 title generation system [17].

These researchers state two advantages of a generative technique over an extractive one. Firstly, a generative approach can create compact representations of text at any compression rate, and secondly they can combine information that is spread across different sentences in the text. However, researchers are now favouring an extractive approach that compresses text using linguistically rich methods because of the difficulty of integrating grammaticality into a generative model of title generation [18]. Nevertheless, generative approaches still have an important role to play in title generation, especially where syntactic information such as punctuation and capitalisation (a prerequisite for most NLP-based techniques) is either missing or unreliable as in the case of automatic speech recognised (ASR) news transcripts.

3 The Topiary Headline Generation System

In this section, we describe the Topiary system developed at the University of Maryland with BBN. As already stated, this system was the top performing headline generation system at DUC 2004. A Topiary-style headline consists of a set of topic labels followed by a compressed version of the lead sentence. Hence, the Topiary system views headline generation as a two-step process: first, create a compressed version of the lead sentence of the source text, and second, find a set of topic descriptors that adequately describe the general topic of the news story. We will now look at each of these steps in more detail.

In [18] Dorr, Zajic and Schwartz stated that when human subjects were asked to write titles by selecting words in order of occurrence in the source text, 86.8% of these headline words occurred in the first sentence of the news story. Based on this result Dorr, Zajic and Schwartz, concluded that compressing the lead sentence was sufficient when generating titles for news stories. Consequently, their DUC 2003 system HedgeTrimmer used linguistically motivated heuristics to remove constituents that could be eliminated from a parse tree representation of the lead sentence without affecting the factual correctness or grammaticality of the sentence. These linguistically motivated trimming rules [3, 18] iteratively remove constituents until a desired sentence compression rate is reached. The compression algorithm begins by removing determiners, time expressions and other low content words. More drastic compression rules are then applied to remove larger constituents of the parse tree until the required headline length is achieved. For the DUC 2004 headline generation task systems were required to produce headlines no longer than 75 bytes i.e. about 10 words. The following worked example helps to illustrate the sentence compression process¹.

Lead Sentence: The U.S. space shuttle Discovery returned home this morning after astronauts successfully ended their 10-day Hubble Space telescope service mission.

Parse: (S (S (NP (NP The U.S. space shuttle) Discovery) (VP returned (NP home) (NP this morning)) (SBAR after (S (NP astronauts) (VP (ADVP successfully) ended (NP their 10-day Hubble Space telescope service mission))))))

1. Choose leftmost S of parse tree and remove all determiners, time expressions and low content units such as quantifiers (e.g. each, many, some), possessive pronouns (e.g. their, ours, hers) and deictics (e.g. this, these, those):

Before: (S (S (NP (NP *The* U.S. space shuttle) Discovery) (VP returned (NP home) (NP *this morning*)) (SBAR after (S (NP astronauts) (VP (ADVP successfully) ended (NP *their* 10-day Hubble Space telescope service mission))))))

After: (S (S (NP (NP U.S. space shuttle) Discovery) (VP returned (NP home)) (SBAR after (S (NP astronauts) (VP (ADVP successfully) ended (NP 10-day Hubble Space telescope service mission))))))

2. The next step iteratively removes constituents until the desired length is reached. In this instance the algorithm will remove the trailing SBAR.

Before: (S (S (NP (NP U.S. space shuttle) Discovery) (VP returned (NP home)) (SBAR after (S (NP astronauts) (VP (ADVP successfully) ended (NP 10-day Hubble Space telescope service mission))))))

After: U.S. space shuttle Discovery returned home

¹ The part of speech tags in the following example are explained as follows: **S** represents a simple declarative clause; **SBAR** represents a clause introduced by a (possibly empty) subordinating conjunction; **NP** is a noun phrase; **VP** is a verb phrase; **ADVP** is an adverb.

Like the ‘trailing SBAR’ rule, the other iterative rules identify and remove non-essential relative clauses and subordinate clauses from the lead sentence. A more detailed description of these rules can be found in [3, 18]. In this example, we can see that after compression the lead sentence reads more like a headline. The readability of the sentence in this case could be further improved by replacing the past tense verb ‘returned’ with its present tense form; however, this refinement is not currently implemented by the Topiary system or by our implementation of this compression algorithm.

As stated earlier, a list of relevant topic words is also concatenated with this compressed sentence resulting in the final headline. The topic labels are generated by the UTD (Unsupervised Topic Discovery) algorithm [3]. This unsupervised information extraction algorithm, creates a short list of useful topic labels by identifying commonly occurring words and phrases in the DUC corpus. So for each document in the corpus it identifies an initial set of important topic names for the document using a modified version of the *tf.idf* metric. Topic models are then created from these topic names using the OnTopic™ software package. The list of topic labels associated with the topic models closest in content to the source document are then added to the beginning of the compressed lead sentence produced in the previous step, resulting in a Topiary-style summary.

One of the problems with this approach is that it will only produce meaningful topic models and labels if they are generated from a corpus containing additional on-topic documents on the news story being summarised. In the next section, we explore two alternative techniques for identifying topic labels, where useful summary words are identified ‘locally’ by analysing the source document rather than ‘globally’ using the entire DUC corpus i.e. the UTD method.

4 LexTrim and TFTrim Headline Generation Systems

In this section, we describe two Topiary-style headline generation systems that use our implementation of the Topiary sentence compression algorithm², but identify pertinent topic labels by analysing the lexical cohesion structure of a news story in the case of the LexTrim system, and term frequency scores in the case of the TFTrim system.

Lexical cohesion is the textual characteristic responsible for making the sentences of a text appear coherent [19]. One method of exploring lexical cohesive relationships between words in a text is to build a set of lexical chains for that text. In this context a lexical chain is a cluster of semantically related proper noun and noun phrases e.g. {boat, ship, vessel, rudder, hull, gallery, Titanic}. These semantic relationships can be identified using a machine-readable thesaurus, in our case the WordNet taxonomy [20]. Here are some examples of these semantic relationships:

- **Synonymy:** *ship* and *vessel* are synonyms because they share the same meaning and can be used interchangeable in text.

² The only significant difference between our compression algorithm and the University of Maryland/BBN approach is that we use Collins’ parser [21], while they use the BBN parser [22].

- **Holonymy:** ship *has part* rudder, therefore ship is a holonym of rudder.
- **Meronymy:** the gallery is *part of* a ship, therefore gallery is a meronym of ship.
- **Hypernymy:** Ship *is a generalisation of* a Titanic, therefore ship is a hypernym of Titanic.
- **Hyponymy:** boat *is a specialisation of* a vessel, therefore boat is a hyponym of vessel.

By clustering semantically related nouns into lexical chains, a more accurate picture of the semantic content of a document can be determined. In particular, lexical cohesion analysis, unlike a term frequency analysis approach, can differentiate between low frequency terms that are ‘genuinely’ unimportant, and low frequency terms that are important topic words because of their strong semantic association with other high content words in the text. For example, in a particular news story, although the noun ‘murder’ occurs only twice in the text, it will be considered an important topic descriptor because of its strong association with terms in a ‘dominant’ lexical chain containing the nouns {homicide, manslaughter, shooting}.

There are three main steps to our technique for identifying topic labels using lexical cohesion analysis. First, the text is processed by a part-of-speech tagger [23], and all proper noun and noun phrases are extracted. These phrases and their location information in the text are then passed as input to the lexical chaining algorithm. The aim of the Chainer is to find relationships between these phrases using the WordNet thesaurus. The Chainer uses a single-pass word clustering algorithm, where the first noun phrase in the news story forms the first lexical chain, and each subsequent phrase is then added to an existing chain if it is semantically related to at least one other noun phrase in that chain. One of the problems with generating lexical chains for news stories is that many of important proper noun phrases will not be present in WordNet since keeping an up-to-date repository of such phrases is a substantial and never ending problem. However, these proper nouns are still useful to the chaining process since they provide an additional means of capturing lexical cohesion in the text through repetition relationships. So our chaining algorithm uses a fuzzy string matching technique to identify full syntactic match (*U.S_President* \Leftrightarrow *U.S_President*), partial full-word match (*U.S_President* \Leftrightarrow *President_Bush*) and a ‘constrained’ form of partial word match between two proper noun phrases (*cave_dwellers* \Leftrightarrow *cavers*). This chaining procedure results in the creation of two distinct sets of lexical chains: WordNet-based noun and proper noun chains, and non-WordNet proper noun chains. A more detailed explanation of our lexical chaining algorithm is given in [24].

The final step, once all lexical chains have been created for a text, is to decide which chain words are the best topic descriptors for the news story. In this way, we can view lexical chaining as a feature extraction method that identifies promising topic labels by virtue of their strength of association with other important noun/proper noun phrases in the text. Noun/proper noun phrase importance, in this context, is calculated with respect to the strength of the lexical chain in which the phrase occurred. More specifically, as shown in Equation 1, the chain strength score is the sum of each strength score assigned to each word pair in the chain.

$$Score(chain) = \sum ((repsi + repsj) * rel(i, j)) \quad (1)$$

where $repsi$ is the frequency of word i in the text, and $rel(i, j)$ is a score assigned based on the strength of the relationship between word i and j . Relationship strengths between chain words are defined as follows: a repetition relationship is assigned a value of 1.0, a synonym relationship a value of 0.9, hypernymy/hyponymy and meronymy/holonymy a value of 0.7. Proper noun chain word scores are assigned depending on the type of match, 1.0 for an exact match, 0.8 for a partial match and 0.7 for a fuzzy match. The lexical cohesion score of a chained word is then the strength score assigned to the chain where the word occurred. These lexical chain words are then concatenated with the compressed lead sentence in order of their lexical cohesion strength, where the number of chain words added depends on the shortfall between the length of the compressed lead sentence and the maximum allowable length of the headline. We have also used this lexical chaining technique to weight the importance of sentence content in an extractive approach to headline generation for closed-caption broadcast news transcripts with segmentation errors; however, no parse-and-trim style sentence compression was employed in that experiment [25].

The third headline generation system examined in this paper, the TFTrim system, employs a much simpler topic labelling strategy, where high frequency words (excluding stopwords) in the news story are added to the topiary-style headline in the order of frequency. In both cases, the LexTrim and TFTrim systems will only assign topic labels that are not included in the compressed sentence part of the headline.

5 Evaluation Methodology and Results

In this section we present the results of our headline generation experiments on the DUC 2004 corpus³. The aim of these experiments was two-fold: to build a linguistically motivated heuristic approach to title generation, and to look at alternative techniques for padding Topiary-style headlines with content words. There are two parts to our evaluation methodology. Firstly, we used the ROUGE evaluation metrics as an automatic means of evaluating headlines, and secondly a randomly selected subset of titles was manually evaluated by a set of human judges. For the DUC 2004 evaluation, participants were asked to generate headlines consisting of no more than 75 bytes for documents on TDT-defined events. The DUC 2004 corpus consists of 625 Associated Press and New York Times newswire documents. The headline-style summaries created by each system were evaluated against a set of human generated (or model) summaries using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics: ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-LCS and ROUGE-W. The first four metrics are based on the

³ Details of our official DUC 2004 headline generation system can be found in [27]. This system returned a list of keywords rather than a 'sentence + keywords' as a headline. It used a decision tree classifier to identify appropriate summary terms in the news story based on a number of linguistic and statistical word features.

average n-gram match between a set of model summaries and the system-generated summary for each document in the corpus. ROUGE-LCS calculated the longest common sub-string between the system summaries and the models, and ROUGE-W is a weighted version of the LCS measure. So for all ROUGE metrics, the higher the ROUGE value the better the performance of the summarisation system, since high ROUGE scores indicate greater overlap between the system summaries and their respective models. Lin and Hovy [2] have shown that these metrics correlated well with human judgements of summary quality, and the summarisation community is now accepting these metrics as a credible and less time-consuming alternative to manual summary evaluation. In the official DUC 2004 evaluation all summary words were stemmed before the ROUGE metrics were calculated; however, stopwords were not removed. No manual evaluation of headlines was performed.

5.1 ROUGE Evaluation Results

Table 1 and Figure 1 show the results of our headline **generation** experiments on the DUC 2004 collection. Seven systems in total took part in this evaluation, three Topiary-style headline generation systems and four baselines:

- The **LexTrim** system, as explained in Section 4, augments condensed lead sentences with high scoring noun phrases that exhibit strong lexical cohesive relationships with other terms in a news story. The **Lex** system is a baseline version of this system, where headlines consist of lexical chain phrases only.
- The **Topiary** system is the University of Maryland/BBN DUC 2004 headline generation system. The **UTD** system, like the Lex system, returns a set of topic descriptors. The UTD algorithm is explained in Section 3. The **Trim** system is another baseline system that only returns the compressed lead sentence as a headline.
- The **TFTrim** system, as explained in Section 4, pads the compressed sentence with high frequency terms found in the original source text when generating a headline. The baseline version of this system is **TF** which returns a sequence of high frequency keywords as the headline.

Table 1. ROUGE scores for headline generation systems on the DUC 2004 collection

	System	ROUGE-1	ROUGE-L	ROUGE-W
Topiary -style systems	TFTrim	<i>0.27933</i>	<i>0.21336</i>	<i>0.12600</i>
	LexTrim	0.25370	0.20099	0.11951
	Topiary	0.24914	0.19951	0.11970
Baseline systems	TF	<i>0.24428</i>	<i>0.17074</i>	<i>0.09805</i>
	Trim	0.20061	0.18248	0.10996
	Lex	0.18224	0.14679	0.08738
	UTD	0.15913	0.13041	0.07797

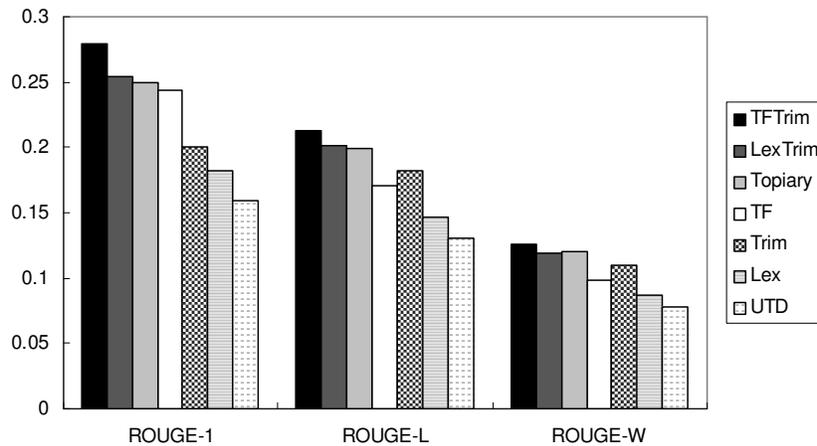


Fig. 1. ROUGE scores for headline generation systems on the DUC 2004 collection

Since the DUC 2004 evaluation, Lin [26] has concluded that certain ROUGE metrics correlate better with human judgements than others depending on the summarisation task being evaluated i.e. single document, headline, or multi-document summarisation. In the case of headline generation, Lin found that ROUGE-1, ROUGE-L and ROUGE-W scores worked best and so only these scores are included in Table 1 and Figure 1. Looking at these scores we can see that the best of the Topiary-style headline systems is the TFTrim system, while the ROUGE scores for the LexTrim and Topiary systems indicate that their performance is very similar. On the other hand, the TF system is the best of the baseline systems where headlines either consisted of a list of keywords (i.e. Lex and UTD) or a compressed sentence (i.e. Trim). Both of these conclusions, suggest that although our lexical chaining method appears to produce better topic descriptors than the UTD method, the best approach is actually the simplest. In other words, the TF technique that uses source document term frequency statistics to identify salient topic labels can outperform both a knowledge-based NLP approach (using WordNet), and a statistical-based approach requiring additional word frequency and cooccurrence information from the entire DUC 2004 corpus.⁴

⁴ In previous headline generation experiments using lexical chains (and no sentence compression), we found that the TF system was outperformed by our gisting system [25]. However, in those experiments we compared sentence extraction rather than word extraction based summarisation. In addition, these experiments were conducted on a broadcast news corpus with segmentation errors, i.e. the end of one news story may be merged with the beginning of the next. We believe that this noise attributed to the poor performance of TF system.

5.2 Manual Evaluation Results

In this section, we report on the results of our manual headline evaluation of the TFTrim and Topiary systems. One of the main criticisms of automatic metrics, such as the ROUGE scores presented in the previous section, is that they do not directly evaluate important summary attributes like readability and grammatical correctness. They also fail to recognise cases where synonymous or semantically similar words are used in the system and reference titles for a news story (e.g. the noun phrase ‘Israeli capital’ is equivalent to the proper noun ‘Jerusalem’), which could result in a system title appearing less relevant than it actually is. It is also unclear whether these ROUGE metrics are sensitive enough to be able to correctly determine the quality of similar style-summaries. To address this problem, we asked four human judges to evaluate the quality of 100 randomly selected headlines generated from the DUC 2004 corpus. These judges were asked to decide, given the human generated titles and the Topiary and TFTrim titles for each document, which system headline was better. In some cases, the system titles were too similar to decide between, so judges were also given a third ‘undecided’ option.

Overall, each of the four judges ranked the TFTrim titles higher than the Topiary titles; however, this result was close with an average of 32.5% of TFTrim headlines and 27.5% of Topiary headlines considered better than the alternative system title. The judges also concluded that 40.0% of titles were too similar to decide between. The average Kappa statistic between each set of human judgements was 0.385 (standard deviation 0.055) which indicates low agreement between judges for this task. One of the factors contributing to this low Kappa score may have been the inclusion of the ‘undecided’ option, as it is obvious from the judgements that judges disagreed most with respect to this aspect of the evaluation. However, even though there is very little difference between the performance of these systems, the aim of these experiments was to determine if Topiary-style summaries require topic descriptors generated from the entire DUC corpus in order to be effective news story headlines. As already stated, one of the problems with the UTD method of topic labelling is that it relies on the existence of a news corpus with similar on-topic documents to the news story being summarised. In many summarisation scenarios such a collection is not readily available, in which case the results of these experiments suggest that keywords identified in the source text are as good as, if not better, than UTD topic descriptors in Topiary-style summaries.

6 Conclusions

In this paper, we have compared the performance of three Topiary-style headline generation systems that use three distinct techniques for ‘padding out’ compressed lead sentences in the automatic generation of news story headlines. The results of our experiment using the ROUGE evaluation suite and a manual evaluation of the system titles, indicate that topic descriptors identified by simple term frequency counts in the source document outperform either keywords identified by a lexical cohesion analysis

of the source text, or statistically derived topic labels from the DUC 2004 corpus using the UTD algorithm.

Following a manual inspection of these system headlines by the authors, it is clear that the strength of the term frequency-based topic labelling method is that it is more consistent in its assignment of quality descriptors to Topiary-style headlines than either of the other labelling techniques. More specifically, the UTD and lexical chaining techniques suffer from the following weaknesses:

- During lexical cohesion analysis, weak descriptors are sometimes chosen from cohesively strong lexical chains. For example, in the case of the following chain {country, Palestine, Israel}, ‘country’ was chosen as an appropriate topic word by virtue of its strong relationship with the other two frequently occurring chain members generated for a particular news story. It is hoped that the inclusion of an *idf* statistic in the lexical cohesion weighting function, described in Section 4, will help to lower the cohesion score of these low content words and improve the performance of the LexTrim system.
- One of the potential strengths of the UTD algorithm is that it can assign topic words to headlines that didn’t occur in the original news story, but are frequently occurring in related on-topic news stories. However, this also commonly leads to the assignment of inappropriate topic labels; for example, in the DUC 2004 corpus there are two prominent topics that frequently mention the country ‘Portugal’, i.e. a topic relating to the former Portuguese colony, East Timor, and a topic discussing the Portuguese Nobel prize winner for Literature, José Saramago. The assignment of the topic label ‘East Timor’ to a headline generated for a news story discussing José Saramago indicates both the dependence of the UTD method on a related corpus of news documents, and the problems associated with the occurrence of related, yet distinct topics in that corpus.

In future work, we intend to proceed by improving both the lexical cohesion score in the LexTrim system, and the sentence compression procedure described in this paper. In addition, we intend to investigate the use of lexical cohesion information as a means of improving the performance of the compression algorithm by helping to limit the elimination of ‘cohesively strong’ parse tree components during sentence compression.

Acknowledgements

The funding support of the Enterprise Ireland is gratefully acknowledged.

References

1. Document Understanding Conference (DUC) <http://duc.nist.gov/>

2. Lin C-Y, Hovy E. Automatic Evaluation of Summaries using n-gram Co-occurrence Statistics. In the Proceedings of HLT/NACCL, 2003.
3. Zajic D., Dorr, B., Schwartz, R. BBN/UMD at DUC-2004: Topiary. In the Proceedings of the Document Understanding Conference (DUC), 2004.
4. Kraaij W., M. Spitters, A. Hulth. Headline extraction based on a combination of uni- and multi-document summarization techniques. In the Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002), 2002.
5. Alfonseca E., P. Rodriguez. Description of the UAM system for generating very short summaries at DUC 2003. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2003), 2003.
6. Copeck T., S. Szpakowicz. Picking phrases, picking sentences. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2003), 2003.
7. Zhou L., E. Hovy. Headline Summarization at ISI. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2003), 2003.
8. Lacatusu F., A. Hickl, S. Harabagiu, L. Nezda. Lite-GISTexter at DUC2004. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2004), 2004.
9. Angheluta R., R. Mitra, X. Jing, M.-F. Moens. K.U. Leuven Summarization System at DUC 2004. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2004), 2004.
10. Alfonseca E., A. Moreno-Sandoval, J. M. Guirao. Description of the UAM System for Generation Very Short Summaries at DUC 2004. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2004), 2004.
11. Kolluru B., H. Christensen, Y. Gotoh. Incremental Feature-based Compaction. In the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2004), 2004.
12. Zhou L., E. Hovy. Template-filtered Headline Summarization. In the Proceedings of the ACL workshop, Text Summarization Branches Out, pp. 56-60, 2004.
13. Witbrock M., V. Mittal, Ultra-Summarisation: A Statistical approach to generating highly condensed non-extractive summaries. In the Proceedings of the ACM-SIGIR, pp. 315-316, 1999.
14. Banko M., V. Mittal, M. Witbrock. Generating Headline-Style Summaries. In the Proceedings of the Association for Computational Linguistics, 2000.
15. Jin R., A.G. Hauptmann. A new probabilistic model for title generation. In the Proceedings of the International Conference on Computational Linguistics, 2002.
16. Berger, A.L., V.O. Mittal. OCELOT: a system for summarizing Web pages. In the Proceedings of the ACM-SIGIR, pp.144-151, 2000.
17. Zajic, D., B. Dorr. Automatic headline generation for newspaper stories. In the Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002), 2002.

18. Dorr B., Zajic D., Schwartz, R. Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In the Proceedings of the Document Understanding Conference (DUC), 2003.
19. Morris J., G. Hirst, *Lexical Cohesion by Thesaural Relations as an Indicator of the Structure of Text*, Computational Linguistics 17(1), 1991.
20. Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Five Papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University, July 1990.
21. Collins M. Three generative lexicalised models for statistical parsing. In the Proceedings of ACL, 1997.
22. Miller, S., M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stones, R. Weischedel. BBN: Description of the SIFT system as used for MUC-7. In the Proceedings of MUC-7, 1998.
23. Xu J., J. Broglio, and W. B. Croft. The design and implementation of a part of speech tagger for English. Technical Report IR-52, University of Massachusetts, Amherst, Center for Intelligent Information Retrieval, 1994.
24. Stokes N. Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking domain. Ph.D. thesis. Department of Computer Science, University College Dublin, 2004.
25. Stokes N., E. Newman, J. Carthy, A. F. Smeaton. Broadcast News Gisting using Lexical Cohesion Analysis. In the Proceedings of the 26th European Conference on Information Retrieval (ECIR-04), pp. 209-222, Sunderland, U.K., 2004.
26. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In the Proceedings of the ACL workshop, Text Summarization Branches Out, pp. 56-60, 2004.
27. Doran W. P., N. Stokes, E. Newman, J. Dunnion, J. Carthy, F. Toolan. News Story Gisting at University College Dublin. In the Proceedings of the Document Understanding Conference (DUC), 2004.