LexTrim: A Lexical Cohesion Based Approach to Parse-and-Trim Style Headline Generation

Ruichao Wang, Nicola Stokes, William Doran, Eamonn Newman,
John Dunnion, and Joe Carthy

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

Abstract. In this paper we compare two parse-and-trim style headline generation systems. 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 cohesion structure of a text. The performance of these systems is evaluated using the ROUGE evaluation suite on the DUC 2004 news stories collection.

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 have already been assigned 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) 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 [1]. The best performing system at this workshop was the Topiary approach [2] which generated headlines by combining a set of topic descriptors generated from the DUC 2004 corpus with a compressed version of the lead sentence, e.g. (Topic Descriptors) *BIN_LADEN EMBASSY BOMBING:* (Compressed Lead Sentence) *FBI agents this week began questioning relatives of the victims.*

Topiary-style summaries perform well in the ROUGE 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 text. 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 two different methods of generated topic labels and observe their effect on summary quality when combined with compressed lead sentences. The Topiary system generates topic descriptors using a statistical approach called Unsupervised Topic Discovery (UTD) [2]. 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 clusters of related documents in order to facilitate the generation of topic models and descriptors, i.e. it needs a structured corpus such as the DUC 2004 collection. 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 additional on-topic news story documents from the DUC 2004 corpus are needed to determine appropriate topic labels for a particular story headline. In Section 2, we describe how we analyse the lexical cohesive structure of news texts using our lexical chaining algorithm. In Section 3, we report on the results of our title generation experiments on the DUC 2004 collection and compare the performance of our system LexTrim with the Topiary approach to this task.

2   The Topiary and LexTrim Headline Generation Systems

In this section, we describe the Topiary system developed at the University of Maryland. This is followed by a description of our headline generation system, LexTrim, which also returns a headline consisting of topic keywords and a compressed version of the lead sentence of the source document. The Topiary system takes a two-step approach to headline generation for news stories:

1. The lead sentence is compressed using the Hedge Trimmer algorithm [2, 3]. This parse-and-trim approach to headline generation removes constituents of a parse tree representing the lead sentence that can be eliminated without affecting the factual correctness or grammaticality of the sentence. A set of linguistically motivated trimming rules is defined in [3] and [2]. These rules iteratively remove constituents until the desired sentence compression rate is reached. Firstly determiners, time expressions and other low content words are removed. More drastic compression rules are then applied to remove larger constituents like trailing prepositional phrases and preposed adjuncts until the desired length is reached.

2. The compressed sentence is then concatenated with a list of relevant topic words generated by the UTD algorithm. This unsupervised information extraction algorithm firstly identifies commonly occurring words and phrases in the DUC corpus. Then for each document in the corpus it identifies an initial set of important topic names 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 model closest in content to the source document is then added to the compressed lead sentence produced in the previous step, resulting in a Topiary-style summary such as the example in Section 1.

The LexTrim system, on the other hand, uses our implementation of the Hedge Trimmer algorithm and a lexical cohesion-based approach to identifying pertinent topic labels. Lexical cohesion is the textual characteristic responsible for making the

sentences of a text appear coherent. 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, yacht, rudder, hull, bow}. The semantic relationships (i.e. synonymy, holonymy, hyponymy, meronymy, hypernymy) are identified using the WordNet taxonomy. Once lexical chains have been generated for a news story, topic phrases are extracted and concatenated with the condensed lead sentence to form a headline. Topic phrases are noun/proper noun phrases that occur in lexical chains with high lexical cohesion scores, i.e. they are phrases that exhibit strong semantic relationships with other important phrases in the text, and so are considered important topic labels. A more detailed description of our lexical chaining algorithm and these cohesion scores can be found in [4].

3   Evaluation Methodology and Results

In this section we present the results of our headline generation experiments on the DUC 2004 corpus using the ROUGE evaluation metrics. In this task, participants were asked to generate very short (<= 75 bytes) summaries of single documents on TDT-defined events. The DUC 2004 corpus consists of 500 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. The format of the evaluation was based on six scoring metrics: ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-LCS and ROUGE-W. The first four metrics are based on the average n-gram match, where $1 \leq n \leq 4$, between a set of model summaries and the system-generated summary for each document in the corpus. ROUGE-LCS calculates the longest common sub-string between the system summaries and the models, and ROUGE-W is a weighted version of the LCS measure. In the official DUC 2004 evaluation all summary words were stemmed before the ROUGE metrics were calculated; however stopwords were not removed.

   Table 1 shows results from our headline generation experiments on the DUC 2004 collection. 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. As explained in Section 3, our approach, LexTrim, augments condensed lead sentences with high scoring noun phrases that exhibit strong lexical cohesive relationships with other candidate terms in a news story. The Lex system in Table 1 returns headlines consisting of lexical chain phrases only. A comparison of the LexTrim, Lex and Trim system results show that the inclusion of lexical chain topic descriptors significantly improves the 'informativeness' of the compressed lead sentence generated by the Trim system.

   Comparing these system results to the performance of the Topiary and UTD DUC 2004 systems, we can see that the Lex system outperforms the UTD system for all ROUGE metrics. This indicates that our lexical chaining method identifies better topic descriptors than the UTD method. A comparison of the Topiary and LexTrim ROUGE scores also indicate that this is the case. This is an interesting result as it shows that a knowledge-based NLP approach (using WordNet) to identifying topic

labels is as good as, if not better than, the statistics-based UTD approach that requires additional word frequency and co-occurrence information from the DUC 2004 corpus before it can predict salient topic labels for a particular document. Hence, our lexical chaining approach is a useful alternative to the UTD method when a corpus containing additional documents on the topic of a particular news story is not available during headline generation.

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

SYSTEM	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-W
LexTrim	0.25370	0.06208	0.02260	0.00870	0.20099	0.11951
Topiary	0.24914	0.06449	0.02122	0.00712	0.19951	0.11891
Trim	0.20061	0.06283	0.02266	0.00792	0.18248	0.10996
Lex	0.18224	0.02903	0.00663	0.00089	0.14676	0.08738
UTD	0.15913	0.01585	0.00087	0.00000	0.13041	0.07797

4   Conclusions

In this paper, we have compared the performance of two headline generation systems that use two distinct techniques for 'padding out' compressed lead sentences in the automatic generation of news story headlines. The results of our experiments using the ROUGE evaluation suite indicate that lexical chain phrases are more informative topic descriptors than statistically-derived topic labels for this task. We intend to proceed in future work by improving the sentence compression procedure described in this paper. Currently, we are investigating the use of lexical cohesion information as a means of improving the performance of the Hedge Trimmer algorithm by limiting the elimination of important parse tree components during sentence compression.

References

1. Lin, C-Y, Hovy E. Automatic Evaluation of Summaries using n-gram Co-occurrence Statistics. In the Proceedings of HLT/NACCL, 2003.
2. Zajic D., Dorr, B., Schwartz, R. BBN/UMD at DUC-2004: Topiary. In the Proceedings of the Document Understanding Conference (DUC), 2004.
3. 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.
4. 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.