A Hybrid Statistical/Linguistic model for Generating News Story Gists

William P. Doran, Nicola Stokes, Eamonn Newman, John Dunnion, Joe Carthy.

Intelligent Information Retrieval Group,
Department of Computer Science,
University College Dublin, Ireland.

{William.Doran, Nicola.Stokes, Eamonn.Newman, John.Dunnion, Joe.Carthy}@ucd.ie

ABSTRACT

In this paper, we describe a News Story Gisting system that generates a 10-word short summary of a news story. This system uses a machine learning technique to combine linguistic, statistical and positional information in order to generate an appropriate summary. We also present the results of an automatic evaluation of this system with respect to the performance of other baseline summarisers using the new ROUGE evaluation metric.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing – *text analysis.*

General Terms

Algorithms, Experimentation.

Keywords

Summarisation; Lexical Cohesion; Machine Learning.

1. INTRODUCTION

Summarisation is a reductive transformation of a source text into a summary text by extraction or generation. It is generally agreed that the optimal solution to automatic summarisation should be based on text understanding that mimics the cognitive processes of humans [6]. However, this level of comprehension by a machine is still a long way in the future. In the interim we must use other properties of a text to try to formulate useful summaries. In this paper we present a machine learning approach to summarisation that combines positional, linguistic and statistical information to automatically generate very short summaries consisting of 10 words or less called gists. More specifically, we use the machine learning algorithm, C5.0 [1], to predict which words in the source text should be included in the resultant gist. We evaluate this system with respect to several baseline gisting systems on a collection of news documents from the DUC (Document Understanding Conference) 2003 corpus [3].

2. SYSTEM OVERVIEW

To create an informative short summary we follow a two-step process: the first step involves creating an intermediate representation of a source text, and the second involves transforming this representation into a summary text. The intermediate representation we have chosen is a set of features, that we feel are good indicators of possible 'summary words'. We focus our efforts on the content words of a document, i.e. the nouns, verbs and adjectives that occur within the document. For each occurrence of a term in a document, we calculate several features: the *tf* or term frequency of the word in the document, the *idf* or inverse document frequency of the term taken from an auxiliary corpus [7], and the relative position of a word with respect to the start of the document in terms of word distance. We also include binary features indicating whether a word is either a noun/verb/adjective, or whether it occurs in a noun or proper noun phrase. The final feature is a lexical cohesion score calculated with the aid of a linguistic technique called lexical chaining. Lexical chaining is a method of clustering words in a document that are semantically similar with the aid of a thesaurus, in our case WordNet. Our chaining method identifies the following word relationship (in order of strength): repetition, synonymy, specialisation and generalisation, and part/whole relationships. Once all lexical chains have been created for a text then a score is assigned to each chained word based on the strength of the chain in which it occurs. More specifically, as shown in the following equation, the chain strength score is the sum of each strength score assigned to each word pair in the chain.

$$Score(chain) = \sum((reps_i + reps_j) * rel(i, j))$$

where $reps_i$ 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. More information on the chaining process and cohesion score can be found in [2, 5].

Then for each word in the DUC 2003 task 1 summarisation corpus (consisting of 624 documents) [3], we assign it a set of values for each of these features, which are then used with a set of gold standard human-generated summaries to train a decision tree summarisation model using the C5.0 machine learning algorithm. The DUC 2003 evaluation provides four human summaries for each document, where words in the source text that occur in these model summaries are considered to be positive training examples, while document words that do not occur in these summaries are considered to be negative examples. Further use is made of these four summaries, where the model is trained to classify a word based on its summarisation potential. More specifically, the appropriateness of a word as a summary term is determined based on the class assigned to it by the decision tree. These classes are

SIGIR'04, July25-29, Sheffield, South Yorkshire, UK.

ordered from strongest to weakest as follows: 'occurs in 4 summaries', 'occurs in 3 summaries', 'occurs in 2 summaries', 'occurs in 1 summary', 'occurs in none of the summaries'. If the classifier predicts that a word will occur in all four of the human-generated summaries, then it is considered to be a more appropriate summary word than a word predicted to occur in only three of the model summaries. This resulted in a total of 103267 training cases, where 5762 instances occurred in one summary, 1791 in two, 1111 in three, 726 in four, and finally 93877 instances were negative. A decision tree classifier was then produced by the C5.0 algorithm based on this training data. To gauge the accuracy of this classifier we used training/test data split of 90%/10%, and found that on this test set the classifier had a precision (true positives divided by true positives and false positives) of 63% and recall (true positives divided by true positives and false negatives) of 20%.

Our gisting technique generates gists consisting of all words that are positively classified by our decision tree summary model, where precedence is given to words depending on their class, i.e. "occurs in 4 summaries" precedes "occurs in 3 summaries" and so on. Preliminary experiments have showed that thresholding of confidence scores for each word class could improve word selection, also we found that some features had a greater impact on the quality of the gist words. We are currently investigating these issues.

In the case where the classifier could not return the required number of words, we then looked at the aggregate feature-weight scores assigned to each word and used the top ranked words according to this score to 'pad out' the gist to its required length. As a post-processing step we identified noun phrase occurrences of these words in the source text. The intuition here was that if we could add more context to a word by representing it in the gist in its original phrasal form then we could improve the readability and quality of the resultant gist. This raises the issue of a trade-off of context versus content, because DUC gists must be no more than 10 words long and as we increase the number of compounds we also reduce the number of content words that the classifier has predicted. To minimise this effect we employ 'severe' pruning heuristics that ensure that only the most salient compound words are added. For example, if we have an adjective-noun compound, we remove the adjective if it cannot be mapped to a noun, e.g. 'magnetic field' will be accepted but 'large field' wont.

3. EVALUATION AND RESULTS

Evaluating summary quality has long been acknowledged as a difficult task. Evaluation can be done either intrinsically or extrinsically; intrinsic evaluation is the preferred method of evaluation as it gauges the quality of the documents directly using human judges. However, an intrinsic evaluation is subjective, expensive and time consuming. To address these problems the DUC 2004 initiative is this year experimenting with a new automatic recall-oriented evaluation metric called ROUGE [4], since it has been shown that this metric correlates well with human summary quality judgements. In essence, the ROUGE metric calculates the n-gram overlap between a system summary and a set of human-generated summaries.

We created several baseline systems to compare the performance of our system. They included TF, which returned the words that correspond to the top ranking *tf* scores, and a RANDOM system that returned words at random from the document. Also we created a LEAD system that returned the content words in the two leading sentence of the document. All of these systems created short summaries of 10-word gists (only content words- nouns, verbs, adjectives) for each of the 624 documents in the DUC 2003 corpus. We then used ROUGE (with the 'stopped' and 'stemmed' switches on) to assess the performance of the systems using the summaries that accompany the corpus. The ROUGE-1 and ROUGE-2 scores are based on the unigram and bigram overlap between the reference and the system summaries, respectively. The results in Table 1 show that our hybrid statistical/linguistic model outperforms all the other baselines on both the ROUGE-1 and ROUGE-2 metrics. The random system is the worst performing gister followed by the Lead Sentence and TF baselines according to the ROUGE-1 score. However, the ROUGE-2 score implies that the Lead Sentence outperforms the TF baseline. Obviously, both high ROUGE-1 and ROUGE-2 scores are preferable; however, from our results we have observed that a trade-off exists between these two measures. Although there are clear benefits of producing gists consisting of phrases rather than words, according to Lin it is the ROUGE-1 score that corresponds most closely with a human-based evaluation [4].

Gisting System	ROUGE-1	ROUGE–2
Hybrid Model	0.3401	0.0633
TF	0.3051	0.0222
Lead Sentence	0.2229	0.0612
Random	0.0750	0.0118

Table 1: ROUGE results for baseline and hybrid gisters.

4. CONCLUSIONS

In this paper we have shown that our hybrid statistical/linguistic gisting technique outperforms a number of baseline gisting systems on the DUC 2003 task 1 portion of the corpus using the ROUGE evaluation metric. A major weakness of these results is that the classifier was tested and trained on the same data. However, initial experiments on sample documents taken from the DUC 2004 corpus indicate that the classifier's performance is consistent with the results reported here.

ACKNOWLEDGMENTS

The support of Enterprise Ireland is gratefully acknowledged.

REFERENCES

[1] Quinlan R. C5.0: *An Informal Tutorial.* RuleQuest, http://www/rulequest.com/see5-unix.html, 1998.

[2] Doran, W., N. Stokes, J. Dunnion, J. Carthy. Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization. In the Proceedings of CICLing, 2004.

[3] DUC 2003: www.nlpir.nist.gov/projects/duc

[4] Lin C-Y, E. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of HLT-NAACL, 2003.

[5] Stokes, N. *Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking domain.* Ph.D. Thesis, Dept. of Computer Science, University College Dublin, 2004.

[6] Spark-Jones, K. *Factorial Summary Evaluation.* Workshop on Text Summarisation in conjunction with SIGIR, 2001.

[7] TDT Pilot Study Corpus: www.nist.gov/speech/tests/tdt