The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space

Ben Hachey, Gabriel Murray & David Reitter
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, UK
bhachey@inf.ed.ac.uk, gabriel.murray@ed.ac.uk, dreitter@inf.ed.ac.uk

Abstract

We present the Embra system, a first-time entry to DUC for 2005 which performed at or above median for the manual assessment of responsiveness and on 4 out of 5 linguistic quality questions. The system takes a novel approach to relevance and redundancy, modeling sentence similarity using a latent semantic space constructed over a very large corpus. We present a simple approach to modeling specificity based on named entities which shows a small improvement over baseline. Finally, we discuss coherence and present a sentence reordering algorithm with a component-level evaluation demonstrating a positive effect.

1 Introduction

DUC 2005 differed from previous years in that it was slightly simplified, with one summarization task rather than several. A central reason for making this simplification was to give researchers a chance to concentrate more on evaluation than on creating new challenges, given that summarization evaluation remains to a large degree an open question. We therefore found this an optimal year to make our first entry into the DUC competition, as there was one straightforward multi-document summarization challenge and a community-wide discussion of evaluation approaches. DUC 2005 investigated both Rouge and Pyramid evaluation schemes in addition to more standard human evaluations of responsiveness and linguistic quality.

The Embra (Edinburgh Multi-document Breviloquence Assay) system is based on a Maximal Marginal Relevance (MMR) framework (Carbonell and Goldstein, 1998), where a single extraction score is derived by combining measures of relevance and redundancy of candidate sentences. The system is novel in that it measures relevance and redundancy using a very large latent semantic space. It addresses specificity by detecting the presence or absence of Named Entities in our extract candidates. And it implements a sentence-ordering algorithm to maximize sentence coherence in our final summaries. This attempts to maximise contextual similarity between the original source document and the summary while also grouping sentences based on similarity in the latent semantic space.

We were encouraged to find that our system performed competitively according to human evaluations, with median or higher scores for responsiveness and for four out of five linguistic quality questions. The fact that our system performed poorly on referential clarity was unsurprising in that our DUC 2005 entry does not address issues such as anaphora resolution or aggregation. In component-level evaluations, we also found slight improvement in Rouge scores when the specificity mechanism is turned on. And in an evaluation which illicited fluency judgments from human readers, we found that the coherence optimization component shows a positive effect.

In the following section, we will briefly overview the preprocessing we carried out. Next, section 3 contains a description of our approaches to relevance and redundancy, specificity, and coherence optimization. Following this, section 4 contains a discussion of our relative performance according to the official human measures produced by NIST. Section 5 contains an analysis of component-level errors. And finally, in section 6, we conclude and discuss future work.

2 Preprocessing

The preprocessing was largely based on LT TTT and LT XML tools (Grover et al., 2000; Thompson et al., 1997) and was adapted from previous work on rhetorical role classification and automatic summarization in the legal domain (Hachey and Grover, 2004). First, we perform to-

kenization, POS tagging and sentence identification. This is followed by lemmatization and named entity recognition.

At the core of preprocessing is the LT TTT program *fsgmatch*, a general purpose transducer which processes an input stream and adds annotations using rules provided in a hand-written grammar file. We also use the statistical combined part-of-speech (POS) tagger and sentence boundary disambiguation module from LT TTT (Mikheev, 1997). Using these tools, we produce an XML markup with paragraph, sentence and word elements having part-of-speech attributes. Further linguistic markup is added using the *morpha* lemmatizer (Minnen et al., 2000) and the *C&C* named entity tagger (Curran and Clark, 2003) trained on the data from MUC-7.

3 System Description

The following three subsections describe the central components of the Embra system for DUC 2005.

3.1 Relevance and Redundancy

A common approach for determining relevance and redundancy in multi-document summarization is to use Maximal Marginal Relevance (MMR), in which candidate sentences are represented as weighted term-frequency vectors which can thus be compared to query vectors to gauge similarity and already-extracted sentence vectors to gauge redundancy, via the cosine of the vector pairs (Carbonell and Goldstein, 1998). While this has proved successful to a degree, the sentences are represented merely according to weighted term frequency in the document, and so two similar sentences stand a chance of not being considered similar if they don't share the same terms. One way to rectify this is to do Latent Semantic Analysis (LSA) on the matrix first before proceeding to implement MMR, but this still only exploits term co-occurrence *within* the documents at hand.

In contrast, our system attempts to derive more robust representations of sentences by building a large semantic space using LSA on a very large corpus. While researchers have used such large semantic spaces to aid in automatically judging the coherence of documents (Foltz et al., 1998; Barzilay and Lapata, 2005), to our knowledge this is a novel technique in summarization.

Using a concatenation of Aquaint and DUC 2005 data (100+ million words), we utilized the Infomap tool[1] to build a semantic model based on latent semantic analysis (LSA) of the corpora. LSA (Landauer et al., 1998) utilizes singular value decomposition of a term/document matrix, with the documents here being newspaper articles. The decomposition and projection of the matrix to a lower-dimensionality space (in this case, 100 dimensions) results in a semantic model based on underlying term relations. There are numerous ways to query the model, such as finding the most closely related words to a given word or deriving a word vector for a given word. Using such word vectors, a given sentence can be represented as a vector which is the average of its constituent word vectors. This sentence representation can subsequently be fed into an MMR-style algorithm. Our implementation of the algorithm (see Table 1) uses λ annealing following (Murray et al., 2005). λ decreases as the summary length increases, thereby emphasizing relevance at the outset but increasingly prioritizing redundancy removal as the process continues.

```
for each sentence in document:
    for each word in sentence:
        get word vector from semantic model
    average word vectors to form sentence vector
    sim1 = cossim(sentence vector, query vector)
    sim2 = highest(cossim(sentence vector, all extracted vectors))
    score = λ*sim1 - (1-λ)*sim2
    if sentence contains multiple named entities:
        if granularity == 'specific':
            weight score higher
        else if granularity == 'general':
            weight score lower
    else:
        do not weight score
extract sentence with highest score
repeat until desired length
```

Table 1: Sentence extraction algorithm

3.2 Specificity

Specificity is addressed in the sentence selection algorithm and is based on the presence of named entities. The intuition behind this is that sentences with more named entities contain specific instantiations of events. The success of our approach also depends on the truth of the converse, i.e. that sentences with fewer named entities contain more generalized event content.

This is currently implemented by boosting the extraction score of a sentence if it contains multiple (two or more) named entities and the granularity is given as specific. If the sentence contains named entities and the granularity is given as general, we down-weight the extraction score. For DUC 2005, we use factors of 1.05 for boosting and 0.95 for down-weighting. These were experimentally chosen through tuning on a small subset of the data.

3.3 Coherence

Work on coherence (or fluency) can be broken down along several dimensions: *discourse coherence*, *cohesion* and *local coherence*. As regards *discourse coherence*, due to constraints of architecture and the sentence extraction framework, the current system is only concerned with telling the story step-by-step in the right order. The

[1] `http://infomap.stanford.edu/`

insertion of discourse connectives can be considered once more reliable techniques are known to detect the discourse structure. Based on shallow cues alone, rhetorical relations can be detected with only around 60 percent accuracy, either with structural and symbolic-based parsing (Marcu, 2000) or trained classifiers (Reitter, 2003).

With respect to *cohesion*, looking at the performance of available, state-of-the-art anaphora resolution algorithms, we decided that it would not be in our interest to substitute pronouns with their (assumed) antecedents. The gain in cohesion would not justify the risk of making factual errors. Pronominalizing full noun phrases would make sense if we could ensure the presence of antecedents, which is rarely the case given the brevity of the summaries.

Local coherence optimizes the transition from one utterance to another, commonly based on the discourse entities that are mentioned in the utterances (a discourse entity is one that can be referred to by a noun phrse). Recent Machine-learning approaches (Barzilay and Lapata, 2005) require knowledge of the entities that are being referred to by each noun phrase in extracted sentences.

In the system, we address discourse coherence by following a number of constraints:

- TIME: Preserve temporal order: mention earlier published events first. Relevant is the date of publication of the original source document for a sentence.

- SEQUENCE: Preserve original presentation order: if two sentences stem from the same document, prefer to present them in their original order.

- CLUSTER: Cluster similar sentences: We use the Cluster 3.0 algorithm to form 2-6 clusters[2], using a standard cosine similarity with the LSA models as similarity function. Sentences in the same cluster are preferably presented together in the target summary.

- CONTEXT: Recreate the original preceding context: Suppose we are to produce a sequence of two sentences $\{A_1|A_2|...|A_n\}B$: we examine the preceding context of B in its original document, and compare it to each candidate target context A_i, selecting the one that bears the highest similarity. This technique has been shown to produce better-than-baseline results in general in multi-document summarization (Okazaki et al., 2004). We use the same LSA model of sentence similarity as for sentence extraction.

[2] `http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm`

These constraints are weighted. We have two sets of weights: a default one, with CLUSTER and CONTEXT weighted strongly, and an alternative one for back-referring sentences, where SEQUENCE is preferred. We consider any sentence that contains anaphors such as *this*, *therefore* as back-referring. DATE is always a weak preference, mainly due to the fact that the timing of described events has only limited bearing on the publication date of a document. The weights were tuned manually, lacking time and data for empirical estimation.

The algorithm is deterministic and optimizes locally: from the bag of extracted sentences, it determines the one that ranks highest with respect to the above contraints and weights, moves it from the bag to the end of the target summary and repeats until all sentences from the bag are inserted. The first sentence is, if possible, a sentence from the extracted set which is also a lead sentence in the original summary, following (Okazaki et al., 2004).

Connectives such as *Thus, ...* were removed using a list of 118 regular expressions, because such connectives only serve a purpose in their original document context. We inserted paragraph breaks between the clusters identified.

4 Official Results

As mentioned earlier, DUC 2005 set a single query-oriented, multi-document summarisation task. There were 50 topic clusters to be summarised with respect to a short topic query consisting of a 1 to 4 sentence description of an information need. An additional constraint indicated whether the summary should be specific or general. There were 31 participating systems. For the results reported here, individual system scores are averaged over topic clusters.

All DUC systems were evaluated manually for responsiveness and five measures of linguistic proficiency. Human evaluation scores for responsiveness (Rsp, defined below), grammaticality (LQ1), non-redundancy (LQ2), referential clarity (LQ3), focus (LQ4), and structure/coherence (LQ5) can be found in Table 2.[3] The Embra system performance is better than mean and median system scores for the responsiveness measure and for three of the five linguistic quality measures (grammaticality, non-redundancy and focus). It is just below mean and median scores for structure/coherence. In terms of referential clarity, the system rank falls to 28 out of 31.

Responsiveness is defined as *the amount of information in the summary that helps to satisfy the information need expressed in the topic*. The fact that the system does fairly well on this measure suggests that the latent seman-

[3] Due to weak Rouge and Pyramid correlations with the responsiveness measure, we focus our discussion here on the human evaluation measures.

	Rsp	LQ1	LQ2	LQ3	LQ4	LQ5
BLine	1.98	4.26	4.68	4.58	4.50	4.00
Min	1.38	2.60	3.96	2.16	2.38	1.60
Mean	2.40	3.76	4.40	2.94	3.11	2.12
StDev	0.30	0.43	0.21	0.43	0.35	0.35
Median	2.44	3.86	4.44	2.98	3.16	2.10
Embra	***2.44***	***3.92***	***4.48***	***2.38***	***3.24***	***2.00***
Max	2.78	4.34	4.74	4.14	3.94	3.24
UpBnd	4.67	4.81	4.91	4.93	4.89	4.76

Table 2: Embra scores compared to average system performance for human metrics.

tic model does a good job of accounting for relevance and redundancy. We leave a proper comparison to standard MMR for future work. One anticipated way to improve on this score in the current sentence extraction framework is to add a sentence simplification module. Besides trimming non-essential information, this should allow more sentences to be included in the summary. LQ4 (focus) and LQ2 (non-redundancy) help confirm that the sentence extraction algorithm is relatively successful.

The system's poor performance in terms of referential clarity is not surprising as there is no model of coreference. Structure/coherence performance is perhaps surprisingly good, on the other hand, given this lack of coreference. We anticipate that the introduction of a module for anaphora resolution will allow significant improvement in both measures. Furthermore, our experience in DUC 2005 has led us to believe that structure/coherence and referential clarity should be considered during sentence extraction.

The baseline system (BLine) in Table 2 was created by taking the first 250 words from the most recent document in the topic cluster. This does very well in terms of the linguistic measures. It is better than all of the submitted systems in terms of referential clarity, focus, and structure/coherence; while for grammaticality and non-redundancy, only a couple of systems perform better than baseline. For responsiveness, however, the baseline does very poorly with only two systems performing worse. The upper bound (UpBnd) is human performance averaged over 4-9 subjects for each cluster. This clearly outperforms all systems on all responsiveness and linguistic quality evaluation measures.

5 Component-level Analysis

5.1 Relevance and Redundancy

An example Embra summary that demonstrates the strength of the LSA extraction method can be seen in cluster d301, one of our highest-rated machine summaries in terms of responsiveness. The query for the cluster concerns organized crime, the countries involved, and the relevant perpetrators. The extracted sentences shown below contain none of the keywords of the query, but are nonetheless clearly relevant. They contain similar words such as *cartel*, *violence*, *assassinations*, *illegally* and *prostitution*.

- The Cali cartel prefers whenever possible to avoid the open violence, including assassinations of high officials, that has focused world attention on the Medellin gangsters.

- In addition to being charged with bringing people illegally into Italy, they were accused of organizing prostitution and providing false documentation.

One of our worst-rated summaries in terms of responsiveness, however, demonstrates the drawback of this extraction approach. The query for cluster d366 regards the commercial applications and potential dangers of cyanide. However, the word *cyanide* never appears in the summary. The sentences may be relevant, but the reader would never know that the subject was cyanide. Even the most naive keyword-spotting extraction approach would have performed better on this cluster.

Extraction could likely be improved by representing sentences differently for measuring redundancy as opposed to measuring query-relevance. This LSA sentence vector representation is suitable for finding sentences relevant to the cluster query, but by using this same representation for measuring redundancy we are likely to reject good candidate sentences simply because there are general underlying similarities between the candidate and already-extracted sentences. Further experimentation will prove if this is the case, but it is hypothesized that a more traditional *tf.idf*-based sentence vector representation will yield improvement in gauging redundancy.

5.2 Specificity

Due to the fact that no evaluation metric addresses specificity explicitly, it is somewhat difficult to analyze the effectiveness of this module. In order to get a rough idea, we compare Rouge scores for the sentence extraction portion of our system with the specificity mechanism switched on and with the specificity mechanism switched off (Table 5.2). We observed an insignificant but positive improvement in the Rouge-2 recall of 0.8% while Rouge-SU4 recall exhibited a slight decrease of 0.2%. For both official DUC 2005 measures, the precision increased giving slightly higher combined F-scores. For Rouge-2, precision improved by 1.2%. And for Rouge-SU4, precision improved by 0.2%.

One problem with this approach is that it is only expected to help specificity by generalizing over the types of events or relations that an entity takes part in (e.g. choosing a sentence such as "The Red Cross provides relief to hurricane victims." for a *general* summary

Rouge		No Spec	NE Spec
2	R	0.05915	0.05965
2	P	0.05972	0.06055
2	F	0.05941	0.06007
SU4	R	0.11758	0.11738
SU4	P	0.11868	0.11891
SU4	F	0.11808	0.11809

Table 3: Comparison of Rouge scores for system with NE specificity (NE Spec) and system with no explicit model of specificity (No Spec).

while choosing a sentence such as "The Red Cross provided food and shelter to victims of Hurricane Hugo in Charleston." for a *specific* summary). This does not however explicitly address conceptual generalization over event types, e.g. 'providing relief' as a supertype of 'providing food and shelter'.

5.3 Coherence

We evaluated whether the coherence module was doing a reasonable job by gathering fluency judgments from two subjects (two of the authors who did not implement the coherence module).

Experiment 1: Subjects were each presented with 10 texts containing only summaries with sentences extracted by the Embra system. Each texts was either re-ordered and optimized for fluency (treatment condition) or randomly ordered with randomly interspersed paragraph breaks (control / baseline condition). Subjects rated random summaries from different document clusters. Each subject rated two texts with the same sentence set stemming from the same document cluster: one in the treatment and one in the control condition. Texts were presented in randomized order.

Subjects were instructed to assign a judgment on a 5-point Likert scale to each sentence in the documents, evaluating the statement

> Perfect coherence: This sentence is either fully related to the previous one, or clearly indicates that it addresses a new topic. The relationship of this sentence to the previous one is clear. It can stand in the given position directly after the previous sentence.

Subjects did not revise choices made in earlier documents in order to avoid a bias introduced by the within-subject experiment design.

Results: An ANOVA by subjects showed that the reordered texts received significantly higher coherence judgements than the scrambled ones ($F(1, 218) = 6.05$, $p < 0.015$). [4] The lower bound (scrambled texts) for our coherence measure is (mean) 2.709, the automatically produced coherent summaries yield 3.155.

Experiment 2: To establish an upper bound, we asked two trained linguists, who remained naïve with respect to the nature of the texts, to rate ten summaries each taken from the set of model summaries, which were handwritten, but fulfilled the same task (query-based, multi-document). *Result:* The mean upper bound was 3.720. This shows that even for humans, creating perfectly coherent summaries is difficult given the emphasis on responding to the given query.

6 Conclusions and Future Work

We have presented the Embra system submitted to DUC 2005. The system is a sentence extraction system which models relevance and redundancy using similarity measures based on a latent semantic space. We presented our approach to building a large semantic space and computing similarity. We presented a simple approach for modeling specificity based on the presence of named entities. And we presented a module for optimizing discourse coherence based on source document context and clustering in the latent semantic space. The overall system performs at median level or better for 4 out of 5 linguistic quality questions and for responsiveness.

We have also presented component-level analysis. We showed that the latent semantic approach to relevance can go wrong when the system chooses sentences that may be relevant but never explicitly state the subject of the query. We also evaluated specificity by using Rouge to measure the system performance with and without this mechanism turned on. This evaluation seems to indicate a small improvement using the NE-based specificity model. Finally, we presented a human evaluation which shows that our discourse coherence approach to sentence reordering performs significantly positive effect. We discussed drawbacks of coherence optimization component being architecturally separate from sentence selection.

To improve the system, we are interested in looking to question answering for methods of treating queries. The current system treats all topics and answers the same. Responsiveness should improve if we do a better job of explicitly modeling question and answer types. Other areas we would like to explore with respect to the sentence extraction module include query expansion and using standard term-based vector similarity for redundancy. And we believe that an explicit evaluation framework for specificity is necessary.

There are various options to increase fluency / coherence in the summaries. Optimizing coherence should

[4] A normalization (z-score) is advised if such Likert-scale judgements are to be used in a different context. For the variance-based correlation tests, however, normalization doesn't make a difference.

mean more than reordering extracted sentences. Coherence, in particular referential clarity, should play a role in the much earlier stage of sentence extraction. Here, choices based on context need to be weighted against the extraction of sentences that address questions in the query. Also, without more reliable anaphora resolution, it will be difficult to optimize cohesion and local coherence. A confidence measure for resolved anaphora may be used to weight sentences during extraction and coherence optimization.

Though not used for the DUC 2005 submission, we also perform chunking, clause recognition, and annotate verbs with features of tense, aspect, voice and modality in the preprocessing stage. We hope to exploit this in future work.

Finally, we are interested in exploring sentence simplification and sentence compression. For preprocessing, we anticipate that sentence simplification will help to isolate the events of interest. Alternatively, in a post-processing stage, we would like to explore sentence compression as a means to trim unnecessary words and include more information. Hovy et al. (2005) discuss attempts to incorporate sentence compression into a summarisation system.

7 Acknowledgments

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI. It was also partly supported by Scottish Enterprise Edinburgh-Stanford Link grant R36410 and, as part of the EASIE project, grant R37588.

We would like to thank the following people for very useful discussion and feedback: James Clarke, Claire Grover, Pei-yun Hsueh, Mirella Lapata, Johanna Moore, and Steve Renals. We would also like to thank Charles Callaway and Peter Bell for taking part in experiments.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, USA.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 2003 Conference on Computational Natural Language Learning*, Edmonton, Canada.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT—a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Ben Hachey and Claire Grover. 2004. A rhetorical status classifier for legal text summarisation. In *Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*, Barcelona, Spain.

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. A be-based multi-document summarizer with sentence compression. In *Proceedings of the ACL-2005 Workshop on Multilingual Summarization Evaluation*, Ann Arbor, MI, USA.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.

Andrei Mikheev. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3).

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.

Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. 2004. Coherent arrangement of sentences extracted from multiple newspaper articles. In *Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence*, Aukland, New Zealand.

David Reitter. 2003. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *LDV-Forum, GLDV-Journal for Computational Linguistics and Language Technology*, 18(1/2).

Henry Thompson, Richard Tobin, David McKelvie, and Chris Brew. 1997. LT XML: Software api and toolkit for xml processing. *http://www.ltg.ed.ac.uk/software/*.