Description of the UAM system at DUC-2005*

Rafael Torralbo, Enrique Alfonseca
Computer Science Department
Universidad Autónoma de Madrid
28049 Madrid (Spain)
Rafael.Torralbo@estudiante.uam.es
Enrique.Alfonseca@ii.uam.es

José María Guirao
Computer Science Department
Universidad de Granada
18071 Granada (Spain)
jmguirao@ugr.es

Antonio Moreno-Sandoval
Department of Linguistics
Universidad Autónoma de Madrid
28049 Madrid (Spain)
sandoval@maria.lllf.uam.es

Abstract

This paper describes the techniques used for our system participating in the Document Understanding Conference 2005. It is a simple sentence-extraction the select the sentences that are more similar to the question provided, using the Vector Space Model.

1 Introduction

The Universidad Autonoma de Madrid has participated in the 2005 Document Understanding Conference. This year, the task consisted in condensing information from 25-50 documents to produce a summary that answers a question. The approach proposed is a simple sentence extraction procedure by calculating the similarity between the collection sentences and the question. Due to lack of time, it was not possible to include special modules for either question-type identification or Named Entity Recognition, which we believe may improve considerably the results.

1.1 Structure of the paper

This paper is structured as follows: Section 2.1 reviews related approaches in multi-document summarisation; Section 3 describes the procedure followed, and Section 4 discusses the conclusions obtained.

2 Related work

2.1 Multidocument summarisation

According to Mani [2001], Multi-Document Summarisation (MDS) systems usually share five steps:

1. **Identification** of the elements to be extracted from the collection.
2. **Matching** instances of these elements across the texts, to find related elements mentioned in different documents.
3. **Filtering** the matched elements, to keep the most salient ones. In this step, the irrelevant elements are filtered out.
4. **Compacting** them, by aggregating and generalising the information from the units that we have kept.

*This work has been sponsored by the grants TIN2004-07588-C03-02 and TIN2004-03140 (Spanish Ministry of Education and Science)

5. **Presenting** the results, for instance, with Natural Language Generation (NLG) or with visualisation methods.

Systems usually differ in the approach chosen for some step.

Unit identification In MDS, it is usual to take each sentence as a single element. All the same, some approaches work with clauses, paragraphs or documents. The compression rate usually affect the kind of unit chosen: if it is very large, then there is generally more need to use units smaller than sentences; and, if it is small, paragraphs can be considered units.

Unit matching A very common practice to match units from different documents consists of using a bag-of-words procedure, by characterising each unit with the set of words contained in it. The vectors can include, together with the words, their frequencies. These can also be transformed into weights, using functions such as tf·idf, χ^2 or Student's t-score. The cosine similarity, calculated by considering the two sets of words as vectors in an N-dimensional space (where N is the size of the vocabulary) is one of the most used similarity functions in Information Retrieval [Salton, 1989]. Other functions used are the scalar product or the Jaccard coefficient. A usual extension to VSM is the dimensionality reduction performed by means of Latent Semantic Analysis (LSA) [Deerwester et al., 1990], which has also been applied in MDS systems [Ando et al., 2000].

Unit filtering Using the similarity metrics from the Vector Space Model, it is possible to cluster the units, so those with a high degree of salient-vocabulary overlapping will be grouped together [Angheluta et al., 2004, Erkan and Radev, 2004, Saggion and Gaizauskas, 2004]. Small clusters with few representatives can also be considered not very important and can be discarded. Some approaches also try to improve the quality of the clusters by filtering the units whose similarity with all the others inside the cluster is not above a threshold [Blair-Goldensohn et al., 2004].

Another procedure is to score the units using a cohesion-based weighting metric [Mani and Bloedorn, 1999]. Possible cohesion relationships are identity relationships between words, synonymy, proximity, coreference and hyperonymy. Sentences whose words have many relationships with words from other units will receive a higher score, and will be selected. Yet another possibility consists in giving higher weights to the NPs that appear in long coreference chains [Witte et al., 2004], and in creating a graph of terms (or events), and next apply a graph-scoring algorithm [Vanderwende et al., 2004, Erkan and Radev, 2004], such as Pagerank [Brin and Page, 1998].

In order to select the most relevant units, there are other heuristics which are also used in *single-document summarisation*, such as unit position and length, or calculating how many terms from the headline appear in every unit. Nobata and Sekine [2004] divide the documents in two groups according to the term distributions, and applies the heuristic based on the unit position just in those groups which appear to contain the key units at the beginning. In MDS, Witte et al. [2004] rank NPs in all the documents based on the length of cross-document coreference chains, and also give highest weights to the NPs that appear in the first units. Finally, the units with the highest-ranking NPs are selected for the summary.

Concerning the size of the units, although most systems mainly work with sentences, some of them also filter clauses and phrases in this step. Common heuristics are to remove relative clauses and appositives [Blair-Goldensohn et al., 2004, Conroy et al., 2004].

Unit compacting If the units have been grouped in clusters in the previous step, it can be expected that the units which are in the same cluster contain repetitive information and, thus, it should be possible to choose just one from each cluster so as to generate the summary. The unit chosen is usually the one closest to the centroid of the cluster [Blair-Goldensohn et al., 2004].

Barzilay et al. [1999] use a more sophisticated approach, by parsing all the units in each cluster with a syntactic analyser, and matching the parse trees with each other. In this matching, they use paraphrasing rules (e.g. transforming passive verbs into active verbs) to discover whether they are *compatible* units. Finally, the units that matched can be merged together with a syntax-based generation procedure.

Results generation and presentation In many cases, systems select units from the documents and put them together. At most, they perform small modifications to them, e.g. by removing relative clauses and appositions, normalising personal names, or removing dangling conjunctions. A few approaches, however, either transform the texts into a logical form [Vanderwende et al., 2004], or apply Information Extraction procedures to fill in templates from the text [Harabagiu and Maiorano, 2002]. In these cases, it is possible to use a Natural Language Generation system to write the summary from the extracted information.

3 The UAM approach

Initially, all the documents to summarise are processed with a pipeline of modules for linguistic processing, using the wraetlic tools version 1.0 [Alfonseca, 2003][1], which include:

- A PoS tagger based on TnT [Brants, 2000].
- A stemmer based on the LaSIE stemmer [Gaizauskas et al., 1995].
- Three chunkers written in C++ and Java, to detect Complex Quantifiers, base Noun Phrases, and complex verbs [Manandhar and Alfonseca, 2000].
- A subject-verb and verb-object detector, written in Java *ad hoc* with hand-crafted rules.

Next, from each document x, the sentences with the highest similarity to the question are selected. Each document is represented as a bag-of-words, containing all the stems for nouns, verbs, adjectives and adverbs. The questions are represented in the same way, and, because they are typically very short, they are also expanded with synonyms using WordNet. Finally, we rank the sentences in order of similarity to the question, using the cosine similarity between the vectors of words. The final summary will be generated from the top ranked sentences. In the previous process, all questions, exclamations and sentences that are clearly incomplete are disregarded.

The next step consists in removing redundant sentences. A similarity metric, again using the Vector Space Model, is calculated between each pair of sentences. Whenever there are two sentences with a very high similarity (in our experiments, 75%), we remove the shortest one.

At this point, we have a set of sentences extracted from different documents. We want to reorder them in the final summary, but we can expect that the relative order of the sentences that come from the same document should be preserved, because that was ordering decided by the journalist. The following algorithm mixes the sentences respecting the relative ordering of the sentences from the same document:

1. Initialise the target summary Sum_k as an empty text.
2. Let $p =$ the first document p_1.
3. Remove the first sentence s from p, and add it at the end of Sum_k.
4. Calculate the similarity between s and the first sentence of all the paragraphs. It is calculated using the VSM, and the similarity metric used is the size of the intersection of the two vectors of words. Note that, now, the first sentence of p is the sentence that was right after s.
5. Let p be the paragraph whose first sentence maximises the similarity, and go back to step (c) with that paragraph. If the best similarity is 0, stop.

Figure 1 shows an example summary obtained with this procedure. It can be seen that, in this case, most of the sentences in the summary have a high degree of word overlapping with the questions: that is the case in which this algorithm will perform better, because it will be easy to identify them.

In contrast, consider Figure 2. In this case, in order to properly answer the question, the system would be able to collect figures of percentage of women in parliaments from different sentences. In other examples,

[1] Available at http://www.ii.uam.es/˜ealfon/eng/research/wraetlic.html

How have relations between Argentina and Great Britain developed since the 1982 war over the Falkland Islands? Have diplomatic, economic, and military relations been restored? Do differences remain over the status of the Falkland Islands?

Today, it is continuing the struggle by economic warfare, rather than by force of arms. Britain has shown growing interest in Latin America as economic reforms across the continent create new trade and investment opportunities. Trade with the UK was growing very quickly, he said. Meanwhile, an Argentine defence ministry official said yesterday that his department began an inquiry of its own last year, in reaction to British investigations, but said it did not get very far. However, after Britain and Argentina established full diplomatic relations last July, Argentine exports grew quickly to a monthly average of about Dollars 20m (Pounds 11.7 m) by the end of last year, against only Dollars 3m previously. Britain is to invite Argentina's economy and foreign ministers to London later this year in the first official visit to Britain by Argentine ministers since the Falklands war in 1982. MRS GILLIAN Shephard, the UK agriculture minister, yesterday stressed the need for co-operation in protecting fragile South Atlantic fish stocks as she prepared to become only the third British cabinet minister to visit Argentina since the Falklands conflict 11 years ago. In December, Argentina announced it would start issuing cut-price licences of its own, breaking the Falklands monopoly. The islanders always knew that Argentina would eventually break their monopoly. British Gas already produces oil in Argentina and last December bought a 29 per cent share in the privatised Buenos Aires gas distribution company. Britain and Argentina re-established diplomatic relations in 1990, a year after President Carlos Menem took office.

Figure 1: Question proposed and answer obtained from the collection of documents.

the answer to a question is an enumeration of countries or people, or data that has to be calculated from information in several documents. In these cases, as it is evident, it is necessary to provide further processing capabilities to the system rather than using a simple sentence extraction procedure. Modules such as the identification of the question type, to know which kind of answer is needed, and a Named Entity Recogniser would probably improve much the results obtained.

4 Conclusions

We present here a system for multi-document summarisation in answer to a question. It is a sentence extraction procedure that chooses the sentence with the highest similarity to the question, using the Vector Space Model. The system currently is not able to answer properly most of the questions in the dataset, as a simple enumeration of sentences is not enough to provide a satisfactory answer. We believe that a module for the identification of question types, combined with a Named Entity Recogniser would greatly improve the results obtained.

References

E. Alfonseca. Wraetlic user guide version 1.0. http://www.eps.uam.es/~ealfon/download.html, 2003.

E. Alfonseca and P. Rodríguez. Description of the UAM system for generating very short summaries at DUC-2003. In *Proceedings of the Document Understanding Conference-2003*, 2003a.

E. Alfonseca and P. Rodríguez. Generating extracts with genetic algorithms. In *Advances in Information Retrieval*, volume 2633 of *Lecture Notes in Computer Science*, pages 511–519. Springer-Verlag, 2003b.

R. K. Ando, B. K. Boguraev, R. J. Byrd, and M. S. Neff. Multi-document summarization by vsualizing topical content. In *Proceedings of the workshop on automatic summarization*, pages 79–88, 2000.

Provide information on numbers of women in parliaments across the world, the gap in political power between the sexes, and efforts that have been made to raise the percentages of women in legislative bodies.

So if there is a second try, we must be sure we can get it through. As a third party with support spread across the country, the first-past-the-post system serves them ill: the Liberal Democrats and their predecessors have never won more than 23 seats (out of up to 650) in a general election since the second world war, despite gaining up to 26 per cent of the vote. In the Senate, the Democrats have, so far, a net gain of one seat. Mr Perot reached 24 per TOTAL 100 29 cent in both Minnesota and Nebraska. However, comparisons between different countries show that some countries which use first-past-the-post do produce proportional results. Labour has already decided on the additional member form of proportional representation for Scotland. Some advance has been made in the representation of women. Some PR systems make it easier to create a more representative legislature. For example, the list system can be used to promote greater representation by women and ethnic minorities simply by ensuring that they appear near the top of the parties lists. Not only do women constitute the majority of the population but they vote in greater numbers-up to 10m more, by some estimates. Many MPs also believe that Ms Jo Richardson, the veteran champion of women's causes, is highly vulnerable, having come bottom in last year's poll with 107 votes. With most primaries for state office still to be held across the country, a Yeakel victory today might well boost the cause of women candidates elsewhere.

Figure 2: Question proposed and answer obtained for a difficult question.

R. Angheluta, R. Mitra, X. Jing, and M.-F. Moens. K. u. leuven summarization system at DUC 2004. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, 1999.

S. Blair-Goldensohn, D. Evans, V. V. Hatzivassiloglou, K. McKeown, A. Nenkova, R. Passonneau, B. Schiffman, A. Schlaikjer, A. Siddharthan, and S. Siegelman. Columbia university at duc-2004. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

T. Brants. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, Seattle, WA, U.S.A, 2000.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

J. M. Conroy, J. D. Schlesinger, J.Goldstein, and D. P. O'Leary. Left-brain/right-brain multi-document summarization. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

H. P. Edmundson. New methods in automatic abstracting. *Journal of the Association for Computational Machinery*, 16(2):264–286, 1969.

G. Erkan and D. R. Radev. The university of michigan at duc 2004. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. University of sheffield: Description of the lasie system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 207–220. Morgan Kauffmann, 1995.

S. Harabagiu and S: Maiorano. Multi-document summarization with GISTEXTER. In *Proceedings of LREC-2002*, 2002.

E. Hovy and C-Y. Lin. Automated text summarization in summarist. In *I. Mani and M. T. Maybury (eds.) Advances in Automatic Text Summarization*, pages 81–94. MIT Press, Cambridge, Massachusetts, 1999.

M. Jaoua and A. Ben Hamadou. Automatic text summarization of scientific articles based on classification of extract's population. In *Proceedings of CICLING-2003*, 2003.

D. Levine. User guide to the pgapack parallel genetic algorithm library, 1996.

C.-Y. Lin and E. Hovy. Identifying topics by position. In *Proceedings of the 5th Applied Natural Language Processing Conference*, pages 283–290, New Brunswick, New Jersey, 1997.

C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the COLING conference*, 2000.

S. Manandhar and E. Alfonseca. Noun phrase chunking with APL2. In *Proceedings of the APL-Berlin-2000 Conference, Berlin. Also published as E. Alfonseca and S. Manandhar, Noun Phrase chunking with APL2, APL Quote Quad (ACM SIGAPL), Vol. 30:4, p. 135-143*, 2000.

I. Mani. *Automatic Summarization.* John Benjamins Publishing Company, 2001.

I. Mani and E. Bloedorn. Summarising similarities and differences among related documents. *Information Retrieval*, 1(1):35–67, 1999.

D. Marcu. Discourse-based sumarization in DUC-2001. In *Proceedings of Document Undestanding Conference, DUC-2001*, 2001.

D. Marcu and L. Gerber. An inquiry into the nature of multidocument abstract. In *Proceedings of the NAACL'01 workshop on text summarisation*, Pittburgh, PA, 2001.

C. Nobata and S. Sekine. CRL/NYU summarization system at DUC-2004. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

J. Otterbacher, A. J. Winkel, and D. R. Radev. The michigan single and multi-document summarizer for duc-2002. In *Document Understanding Conference, DUC-2002*, 2002.

H. Saggion and R. Gaizauskas. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

G. Salton. *Automatic text processing.* Addison-Wesley, 1989.

L. Vanderwende, M. Banko, and A. Menezes. Event-centric summary generation. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.

R. Witte, A. Bergler, Z. Li, and M. Khalifé. Multi-erss and erss 2004. In *Proceedings of the DUC-2004 Workshop*, Boston, MA, 2004.