Googling answers' models in question-focused summarisation*

Enrique Alfonseca[1,2] **Manabu Okumura**[2] **José María Guirao**[3] **Antonio Moreno-Sandoval**[4]

[1]Computer Science Dept. Univ. Autónoma de Madrid Enrique.Alfonseca@uam.es

[2]Precision and Intelligence Laboratory Tokyo Institute of Technology oku@pi.titech.ac.jp enrique@lr.pi.titech.ac.jp

[3]Computer Science Dept. Univ. de Granada jmguirao@ugr.es

[4]Department of Linguistics Univ. Autónoma de Madrid sandoval@maria.lllf.uam.es

Abstract

This paper describes the techniques used for our system participating in the Document Understanding Conference 2006. We describe a new system, built from scratch, that focuses primarily on collecting models of possible answers for each question from the Internet, and applying those models to select the answer sentences from the documents in the collection. The system performed averagely in the manual evaluation done by NIST.

1 Introduction

The DUC-2006 task was basically the same as in DUC-2005 (Dang, 2005): to synthesise, from a set of 25–50 documents, an answer to a non-factual question. In this paper, we describe the procedure chosen for approaching this task.

As pointed by some authors in DUC-2005 (Blair-Goldensohn, 2005; Lacatusu et al., 2005), an analysis of the questions, although it is not necessarily essential in producing good summaries, is possibly a great help in guiding the system towards answering the user's information need. Secondly, the expansion of question words using either contextual co-occurrences, Latent Semantic Analysis (Landauer and Dumais, 1997; Hachey et al., 2005) or synonyms (Hovy et al., 2005) is already widely-used in text summarisation and related fields. Therefore, we have decided to try both approaches in a slightly modified way: the queries are analysed to discover which are the particular user's information needs and which question terms refer to the domain being addressed; and the questions are expanded by searching the web for similar documents, so a medium-size corpus is collected for each user need.

The following section describes the approach followed, and Section 3 discusses the results obtained and open lines for future work.

2 System overview

The procedure for summarising a document set has three consecutive steps:

1. Questions processing.
2. Collection of models for possible answers to each question.
3. Selection of sentences from the documents to be summarised.

The following subsections elaborate each step.

2.1 Analysis of questions

In the first step, all the questions are analysed to discover which should be answered for each of them.

Initially, all the questions are pre-processed with a pipeline of modules for linguistic processing, using the wraetlic tools version 2.0 (Alfonseca et al., 2006), which includes modules for tokenisation, sentence splitting, part-of-speech tagging, lemmatising, Named Entity identification and partial parsing. Right after this step, using the output of the parser, whenever there are several questions co-ordinated with a conjunction the program splits them into separate sentences. So, for instance, questions such as

(1) What happened and how should it be solved?

*This work has been partially supported by the grants TIN2004-03140 and TIN2004-07588-C03-02

Dataset	D0601A
Title	Native American Reservation System - pros and cons
Question 1	Discuss conditions on American Indian reservations or among Native American communities
Verb	Discuss
Theme	[conditions]
Background terms	[American_Indian_reservation, Native_American_community]
Question 2	Include the benefits and drawbacks of the reservation system
Verb	Include
Theme	[benefits, drawbacks]
Background	[reservation_system]
Question 3	Include legal privileges and problems
Verb	Include
Theme	[legal_privileges, problems]
Background	[]

Table 1: Output of the question processing step for the first document set in the test corpus.

will be divided into the following two:

(2) a. What happened?

b. How should it be solved?

All questions contain, at the beginning, either a wh-word (e.g. *who,what,how...*) or an imperative verb (e.g. *Discuss,Include*). Depending on the case, the system extracts the following information:

- The *wh-word* or the *verb* that are introducing the question. If there are both, the *wh-word* only is provided.
- The theme of the question. This is collected in the following way:
  - If the question started with a verb, the head of the NP that is direct object of that verb is considered the question's *theme* that should be answered.
  - Otherwise, depending on the kind of wh-word, either the head of the NP that contains the wh-word or the head of the next NP is collected as *theme*. A particular case is the wh-word *why*, for which the theme is set to *reason, cause* regardless of the remainder of the question.
- Finally, all the other words are collected as background terms for that question.

Both for the themes and for the background terms, words belonging to basic Noun Phrases are returned together, as multi-words. Table 1 shows the output of the question analysis step for the first document set in the test corpus. There are three questions, from which the system discovers that it is necessary to answer five different themes: conditions, benefits, drawbacks, legal privileges and problems. The background terms indicate that the topic concerns the Indian reservations.

2.2 Googling answer's models

In the second step, for each question theme, a text collection is downloaded from the Internet to construct a word model about it. The procedure consists of the following steps:

1. A query is sent to the Google search engine, including the theme as a compulsory keyword, and the background terms as optional keywords. In general, the background terms from all the questions in the same dataset refer to the same topic, so the query for a given question is completed with the background terms from previous questions.

   For instance, from the example in Table 1, the following five queries will be sent:

```
conditions AND
(American_Indian_reservation
OR Native_American_community
OR reservation_system)

benefits AND
(reservation_system OR
American_Indian_reservation
OR Native_American_community)
```

condition	108.51	**benefit**	107.01	**drawback**	145.82	**privilege**	230.61	**problem**	68.59
reservation	43.18	offer	38.00	discusses	9.41	**legal**	230.61	indian	11.22
life	33.39	prison	28.50	role	9.41	non-white	41.93	american	7.57
weather	16.69	spirituality	19.00	reality	9.41	exemption	41.93	native	7.48
total	16.69	don	19.00	group	9.41	specifically	41.93	united	7.42
subjectivity	16.69	practice	19.00	time	9.41	nature	20.97	alaska	5.56
care	16.69	university	19.00	virtual	9.41	slave	20.97	committee	5.56
hpwren	16.69	also	19.00	business	9.41	appropriate	20.97	face	5.56
live	16.69	gaming	19.00	just	9.41	dial	20.97	hawaiian	5.56
way	16.69	potential	10.50	well	5.22	special	20.97	unique	5.56
chronic	9.02	t	10.50	use	5.22	grant	20.97	phoenix	3.71
school	9.02	center	10.50	operation	4.70	anachronism	20.97	type	3.71
economic	9.02	vacation	9.50	age	4.70	longer	20.97	continued	3.71
improve	8.35	digest	9.50	usually	4.70	direct	20.97	history	3.71
home	8.35	anaffidavit	9.50	blather	4.70	court	20.97	affairs	3.71
conditionsof	8.35	process	9.50	issue	4.70	western	20.97	stem	3.71
housing	8.35	streamlined	9.50	datum	4.70	charter	20.97	states	3.71
return	8.35	andarea	9.50	satisfying	4.70	colonial	20.97	jurisdiction	3.71
candidate	8.35	prisoner	9.50	context	4.70	status	20.97	medical	3.71

Table 2: Top ranking words in the χ^2 vectors for the five themes to be answered in document set D0601A.

```
drawbacks AND
(reservation_system OR
American_Indian_reservation
OR Native_American_community)

legal_privileges AND
(reservation_system OR
American_Indian_reservation
OR Native_American_community)

problems AND
(reservation_system OR
American_Indian_reservation
OR Native_American_community)
```

For each query, one thousand document snippets (at most) is downloaded from the web.

2. Next, each of the snippets are modelled as vectors of words and co-occurrence frequencies. At this step, several sizes of context have been tested, such as taking the whole snippets, or considering as context just the sentences containing the theme words. Finally, the sentence-based context size was considered the best one based on a manual observation of the summaries generated for the training set.

3. Now, we would like to keep just the words that are representative for the topic that is being answered. Therefore, a χ^2 test is performed against the British National Corpus so that only words with χ^2 values over 0 are retained.

4. At this point, we have a vector-space model of the main topic of the questions, and we would need to differentiate the model for each of them so we can answer each question separately in the output summary. Therefore, the signatures are now contrasted to each other. Three different weight functions have been applied: χ^2, *tf·idf* and log likelihood (Dunning, 1993), and the best results (again, based on a manual observation) have been obtained with a combination of χ^2 and *tf·idf*.

After the second step, for each one of the themes to be answered there is a separate vector that contains some relevant words. Table 2 shows the signatures obtained, using the χ^2 weight function. As can be seen, the vector for conditions includes relevant terms such as *housing, economic* or *school*. *University* and *gaming* are listed as benefits of the reservation system, and *exception* or *grant* as legal privileges.

2.3 Summary generation

In order to choose the best sentences, the cosine similarity is calculated between each of the sentences in the original documents and the model vectors. Because we have vectors for several weight functions, this provides us with several separate rankings of sentences for each of the themes. Furthermore, a sentence may be shared between rankings belonging to different themes.

Now, a single ranking is calculated for each theme in the following way: for each sentence, its position in all the rankings is added up, and they

Concerning **conditions** , with economic opportunities on reservations lagging behind those available in big cities , and with the unemployment rate among Native Americans at three times the national average , thousands of poor , often unskilled Native Americans are rushing off their reservations. New homes are being built on the reservation by retirees who , after living in black society in cities like Washington and Los Angeles , have returned to the reservation , where they pay no property taxes.
Concerning the **benefits and drawbacks** , frustrated by years of distressing results , schools and groups like the National Indian Education Association have begun pressing states and the federal government for more money for academics and crumbling buildings , programs to train Indian teachers , and support for parents whose poverty , substance abuse or unemployment leave them unmotivated or unable to help their children stay in school and achieve.
Concerning **legal privileges and problems** , a court should ever rule in the Shinnecock's favor , the tribe would hold legal title to billions of dollars in property. In Boston , at Dunne's Boston Indian Council , attendance lags because it does not provide substance abuse treatment , legal services or on-site training programs. Expensive problems that keep people out of work , such as drug and alcohol addiction , domestic violence , poor education and teen births , and lack of jobs , are more prevelent in Indian communities , he noted. Federal programs distributed to American Indians based on census data include the Native American Employment and Training Programs , grants to local education agencies for Indian education , and family violence prevention and services.

Figure 1: Summary generated for the first set in the test corpus.

are sorted according to that value. Thus, the higher a sentence is positioned in all rankings, the higher is is placed in the unified one.

Next, for each of the themes, a mini-summary is generated taking the top n sentences, with the conditions that:

- A sentence is not selected for a theme if it had also been selected for other theme in whose ranking it is positioned higher.
- A sentence is not selected for a theme if another sentence with a high overlapping has been selected for other theme in whose ranking it is higher.

Finally, a small introduction is written at the beginning of the answer for each question, consisting of the word *Concerning* followed by the theme name, aiming at giving the summary a little more coherence. Figure 1 shows the final summary generated for the data set in the example. Note that the answers to several themes are put together if they belong to the same question.

3 Discussion

We describe here out contribution to DUC-2006. The main novelty of this approach is the automatic collection of answer's models from the web that will be used later in selecting the best sentences and generating the summary.

In our settings we set the condition that the lengths of all the mini-summaries together must not exceed the total limit of 250 words. However, in our system, punctuation was not taken into consideration in this count, although it was printed separately from the words. It seems that, in the NIST evaluation, commas have been considered separate words, and thence most of our summaries have been truncated to around 230-240 words.

According to the averaged results provided by NIST, our system, with id. 30, ranked in the 15th position for overall responsiveness, in the 17th position for content responsiveness, and in the 11th position in linguistic quality (out of 35 systems), being in middle positions in the table in all cases. We consider this a good result considering the simplicity of the sentence ranking and summary generation procedure applied, and the evaluation problem that truncated most of our summaries.

Unfortunately, due to our time restrictions we could not participate in the pyramid evaluation.

We believe that the use of the web to automatically generate answers' models seems a promising line of research if combined with more sophisticated sentence selection and reordering strategies. Possible improvements we have in mind are including an anaphora resolution module for personal pronouns, and selecting the sentences with Maximal Marginal Relevance (MMR) (Carbonell

and Goldstein, 1998).

References

E. Alfonseca, A. Moreno-Sandoval, J. M. Guirao, and M. Ruiz-Casado. 2006. The wraetlic NLP suite. In *Proceedings of the Language Resources and Evaluation Conference, LREC-2006*, Italy.

S. Blair-Goldensohn. 2005. From definitions to complex topics: Columbia university at DUC-2005. In *Proceedings of the Document Understanding Conference-2005*.

J. G. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne, Australia. ACM.

H. T. Dang. 2005. Overview of duc-2005. In *Proceedings of the Document Understanding Conference-2005*.

T. E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

B. Hachey, G. Murray, and D. Reitter. 2005. The embra system at DUC 2005: Query-oriented multi-document summarization with a very large latent semantic space. In *Proceedings of the Document Understanding Conference-2005*.

E. Hovy, C.-Y. Lin, and L. Zhou. 2005. A BE-based multi-document summarizer with query interpretation. In *Proceedings of the Document Understanding Conference-2005*.

F. Lacatusu, A. Hickl, P. Aarseth, and L. Taylor. 2005. Lite-GISTexter at DUC 2005. In *Proceedings of the Document Understanding Conference-2005*.

T. K. Landauer and S. T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.