Evaluation of a Cross-lingual Romanian-English Multi-document Summariser

Constantin Orăsan, Oana Andreea Chiorean

Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Wolverhampton, United Kingdom
C.Orasan@wlv.ac.uk

Abstract

The rapid growth of the Internet means that more information is available than ever before. Multilingual multi-document summarisation offers a way to access this information even when it is not in a language spoken by the reader by extracting the gist from related documents and translating it automatically. This paper presents an experiment in which Maximal Marginal Relevance (MMR), a well known multi-document summarisation method, is used to produce summaries from Romanian news articles. A task-based evaluation performed on both the original summaries and on their automatically translated versions reveals that they still contain a significant portion of the important information from the original texts. However, direct evaluation of the automatically translated summaries shows that they are not very legible and this can put off some readers who want to find out more about a topic.

1. Introduction

Even though a large number of newspapers make their information available on their websites, it is still very difficult to know what is happening in different parts of the world unless the events are dramatic enough to capture the attention of the world media. There are two main reasons for this: Quite often the news is not in a language understood by the reader therefore making it impossible to read. Even in the cases where the language does not constitute a barrier, quite often the amount of information available is so large that it is impossible to read everything published. A solution to this problem is offered by *multi-document multilingual summarisation*, a branch of summarisation which produces summaries from several documents and employs techniques from automatic translation to generate an output in a language other than the language of the input (Mani, 2001).

This paper presents a system which facilitates access of English speakers to Romanian news by employing a multi-document summariser, whose results are automatically translated into English using a freely available Romanian to English translation engine. By using this approach, we hypothesise that it is possible to give readers access to the most important information in the texts. In order to confirm or dismiss this hypothesis, judges are asked to answer multiple choice questions on the basis of automatic summaries in Romanian and their automatically translated versions in English.

The structure of the paper is as follows: The next section briefly presents some background information about work in multi-document and multilingual summarisation. Section 3 describes the summarisation method employed to summarise Romanian texts, followed by its evaluation in Section 4. Section 5 evaluates the automatically translated summaries and discusses whether they constitute a good way of accessing information. The paper finishes with conclusions.

2. Related work

Due to the increase in the amount of available information it is no longer enough to summarise single documents, and it is necessary to be able to produce summaries from collections of documents on related topics. As a result of dealing with multiple documents, researchers have to face a greater number of challenges than in single document summarisation, challenges which include occurrence of more redundant information, contradictory information, mis-ordering of events, etc. This section briefly presents work related to multi-document and multilingual summarisation. More detailed information about these topics can be found in Mani (2001), Hovy (2003) or Sparck Jones (2007).

One of the most serious challenges that need to be addressed in multi-document summarisation is the occurrence of redundant information. Maximal Marginal Relevance (MMR) (Goldstein et al., 2000) is a method that identifies sentences relevant to a query, while trying to reduce the repeated information. Radev et al. (2000) and Radev et al. (2001) treat automatic summarisation as a clustering problem and extracts the centers of the identified clusters. A method inspired by the automatic generation of hypertext was proposed in (Salton et al., 1997) and successfully used to produce summaries of single and multiple documents. Barzilay et al. (1999) approach the multi-document summarisation task from the text generation perspective. They find common phrases in documents related to an event and use them as input for a language generation system. Similarities and differences between text units are identified in (Mani and Bloedorn, 1999) by building a graph representation of the document.

The domain of multi-document summarisation is currently a very active research area as a result of the Document Understanding Conferences (DUC)[1] and the forthcoming Text Analysis Conference[2]. Both conferences feature tasks where users need to produce various types of summaries

[1] http://duc.nist.gov/

[2] http://www.nist.gov/tac/

from multiple documents.

Multilingual summarisation is more difficult than multi-document summarisation due to the fact that it also involves some form of automatic translation. As a result, less work has been reported in this field. A large evaluation experiment for English and Chinese multilingual summarisation is presented in (Radev et al., 2002). The Multilingual Summarization Evaluations (MSE)[3] organised in 2005 and 2006 also addressed the problem of multilingual and multi-document summarisation. The training data used in these evaluations contains 25 clusters from DUC2004. These clusters comprise news stories in English and Arabic, as well as automatic translations of Arabic texts into English.

3. Multi-document summarisation for Romanian

The previous section mentioned several multi-document summarisation methods. This section describes the approach employed in this paper and how it was adapted to process Romanian texts. We first start the section with an explanation of how we extracted clusters of documents related to a topic selected by the user. The second part of this section explains how these clusters were processed in order to produce summaries.

3.1. The clusters

Multi-document summarisation methods are normally applied to clusters of documents linked to a user topic which is usually expressed as a list of keywords for a retrieval engine. In this research, the related documents are retrieved using ht://Dig, a search engine suitable for use with finite collections, such as intranets or local computers.[4] The query used to retrieve documents is formed from the words which identify the topic of interest, and the retrieved documents are required to contain all these words. Due to the fact that ht://Dig cannot be normally used to search the Internet, the collection of articles which was used here had to be downloaded onto our computers first. ht://Dig functions very much like a search engine, and therefore it can be easily replaced by any search engine which retrieves documents from the web.

Given that the retrieval engine runs on our computer, we were able to adjust its parameters to best fit our requirements. In order to limit the computation necessary for producing summaries, we decided not to process the entire retrieved documents, but instead to process only snippets returned by the search engine from these documents. We restricted these snippets to 10,000 characters including the white spaces and we considered only the first 50 snippets retrieved by the ht://Dig.[5] If in the future we decide to change the retrieval engine, it may be necessary to process the full documents because quite often the snippets returned by search engines such as Google are too short to allow proper processing.

The retrieved snippets are fed into the summarisation module which extracts from them the most pertinent sentences to be included in a summary.

3.2. The summarisation method

As shown in Section 2., most of the methods used in multi-document summarisation were initially proposed for English, but some of them can be easily adapted to other languages. The method used here to produce summaries from clusters of related documents relies on the Maximal Marginal Relevance (MMR) method proposed in (Goldstein et al., 2000). The reason for choosing this method was that it requires little language dependent information, and that the only linguistic preprocessing it needs is tokenisation and sentence segmentation. Other methods were dismissed because they require linguistic resources which are not available for Romanian (e.g. parsers, dictionaries of synonyms, etc.). In addition, the method is genre independent so it can be easily applied to any types of documents and it is very fast which means it can process large collections of documents in short time.

The MMR method extracts sentences from a text on the basis of scores assigned to them by a formula that tries to maximise the similarity of the selected sentences to the user topic and minimise the redundant information in the summary. A summary is produced through an iterative process in which the sentence with the highest score is added to the summary and the score of sentences not extracted yet is recalculated. This process is repeated until the desired length is reached. The formula used here to score a sentence is:

$$MMR(Q,R,S) = \arg \max_{D_i \in R \setminus S} (\lambda * sim_1(D_i, Q) - (1-\lambda) * \max_{D_j \in R} sim_2(D_i, D_j))$$

where Q is the user topic used to produce the summaries, R is the set of sentences retrieved on the basis of Q and S is the set of sentences extracted so far. The sim_1 calculates the similarity between a sentence D_i and the user query Q, whilst sim_2 is an anti-redundancy metric which calculates how much of the information in D_i is already present in the extracted sentences S. The λ parameter has a value between 0 and 1, and offers a way to balance the amount of new information to be added to a summary with how similar the already extracted information should be to the user query.

The similarity between sentences is calculated using cosine similarity (Manning and Schütze, 1999) between tokens contained in the sentences. In order to have a better idea about the performance of the system, two types of tokens were considered: words as they appears in texts and words truncated to 5 and 6 characters. Each token was weighted using TF*IDF (Salton and McGill, 1983) before it was used to calculate the similarities. A demo of the multi-document summarisation system is available at http://clg.wlv.ac.uk/demos/ro-mds/.

[3] http://research.microsoft.com/ lucyv/MSE2006.htm

[4] More information about ht://Dig can be found at http://www.htdig.org

[5] It should be pointed out that in not all the cases the retrieval engine returned more than 50 snippets and only some of the snippets had 10,000 characters.

4. Evaluation

The aim of this research is to see whether fully automatic machine translation combined with multi-document summarisation can facilitate access to information in a language not known by the reader. However, before assessing this, it is necessary to determine the performance of the summarisation method employed. This section evaluates the method used to produce the summaries in Romanian.

4.1. Settings for the evaluation

In order to evaluate the summarisation method described in the previous section, a corpus of Romanian newspaper articles published between 2001 and 2005 was built. From this corpus, five topics were selected for evaluation. These topics referred to important events which affected Romania and which were covered extensively in the media, but are little known outside Romania. Table 1 presents the five topics.

Evaluation of automatic summarisation is a difficult process due to the fact that there is not only one 'perfect abstract' which should be matched by the machine, but a multitude of summaries which are perfectly acceptable. From the existing evaluation methods available, we decided to use a task-based evaluation, where human judges had to answer multiple choice questions on the basis of a text they were given. Multiple choice questions about the most important information from each cluster were produced prior generating any automatic summaries. Some of these questions had *Yes/No* answers, whilst others had several possible answers. In the latter case, close distractors were introduced in order to check the validity of the answer. For both types of questions, an additional answer *I don't know* was introduced for those cases where the judges could not decide which was the correct answer on the basis of the text they were given. An example of Yes/No question is:

> *Is NATO interested in establishing military bases in Romania?*
>
> - *Yes*
> - *No*
> - *I don't know*

An example of question with several options is:

> *With which parties is Basescu hoping to achieve a parliamentary majority?*
>
> - *PUR and UDMR*
> - *PSD and PRM*
> - *PUR and PSD*
> - *UDMR and PSD*
> - *I don't know*

The difficulty of this question is that all the abbreviations designate political parties that planned to be part of the government, so it is difficult for the judges to guess the answer. The quality of a summary was measured by the number of questions which could be answered correctly on the basis of the summary.

Automatic summaries were produced using four different sets of parameters:

- MMR1: token = truncation to 6 characters, $\lambda = 0.5$
- MMR2: token = word, $\lambda = 0.5$
- MMR3: token = truncation to 6 characters, $\lambda = 0.6$
- MMR4: token = truncation to 5 characters, $\lambda = 0.6$

These particular sets of parameters were selected for evaluation because empirical observation of the results indicated that they lead to good results. In all cases a stoplist was used to filter out stopwords, TF*IDF was employed to weight the tokens and the produced summaries had around 2000 characters including whitespaces.

In addition to these four methods, summaries produced by a baseline method and human written summaries were evaluated. The baseline method extracted the first sentence of each of the retrieved articles until the desired length was reached. The extracted sentences were ordered by the date on which they were published. The decision to employ this baseline relies on the fact that quite often the first sentences of newswire texts produce a good summary of the text.

The human summaries were produced in order to find an upper limit of our summarisation methods. Due to time and resources, only one summary per topic was produced using an extractive approach (i.e. sentences were extracted from the clusters, but were not assembled in a coherent abstract). In order to produce these summaries, the human summariser was given the snippets retrieved by ht://Dig and asked to extract those sentences which were the most important to the given topics until the target length was reached. No other instructions were given. It should be pointed out that different persons produced the human summaries and the questions used in the evaluation.

4.2. The evaluation results

For the evaluation, human judges were given summaries and asked to answer questions on their basis. Each judge was shown only one summary of a topic so that the answer to a question was not influenced by information the judge had seen before in another summary. In addition, the judges were asked to answer only on the basis of information present in the summary, and not on the basis of their knowledge about the events. For this experiment, we managed to find 60 judges so that 10 different people evaluated the results of each method. The percentages of correctly answered questions are presented in Table 2.

As expected, the highest number of questions correctly answered was noticed when the judges received human written summaries, whereas the lowest one when they received the baseline summaries. For the automatic summarisation method the best results are obtained by the MMR3 method which used the following parameters: truncation to 6 characters, $\lambda = 0.6$, stoplist and TF*IDF. To our surprise, there were cases where the human produced summaries contained less answers than some of

1	ARDAF wants to pay to stop Petrovschi scandal
2	Basescu forms the government with UDMR and PUR
3	American bases in Romania
4	Flat-tax rate from 1st of January 2005
5	Romanian journalists kidnaped in Iraq

Table 1: The selected topics

	Human	Baseline	MMR1	MMR2	MMR3	MMR4
Topic 1	84%	14%	46%	46%	52%	0%
Topic 2	48%	10%	62%	40%	50%	50%
Topic 3	60%	40%	42%	48%	74%	64%
Topic 4	58%	48%	64%	74%	78%	64%
Topic 5	60%	38%	42%	37%	33%	42%
All topics	62%	30%	51%	48%	57%	44%

Table 2: The results of the task-based evaluation on the Romanian summaries

the automatic summaries. The explanation for this is that the person who wrote the summaries considered topics other than those covered by the questions as important.

4.3. Evaluation of the coherence

In addition to asking judges to answer questions on the basis of texts given to them, they were also asked to rate the coherence of each summary on a scale from 1 to 5. Table 3 presents the average scores obtained by different types of summaries.

As can be seen in the table, the human summaries obtain the highest score despite the fact that the person who produced them did not produce coherent texts deliberately. The explanation for this is that the human summariser chose a certain set of events from the cluster as important and selected sentences linked to that event. In this way, the sentences connect much better than those in the automatic summaries. The baseline features the lowest cohesion score, whereas summaries produced by MMR1 were ranked as the most coherent ones, followed by those produced by MMR3. These results are very similar to those observed in Table 2, the only difference is that summaries produced with MMR3 can be used to answer more questions than those produced with MMR1.

In this section, our implementation of the MMR summarisation methods was evaluated and the combination of parameters which leads to the best summaries has been identified. The evaluation also revealed that the percentage of questions correctly answered using the best summarisation method is not much lower than the percentage of questions answered using the human summaries. For this reason, it can be concluded that the summarisation method employed contains enough information to be used instead of the human produced ones.

5. Automatic translation as a means of accessing Romanian news stories

The evaluation described in the previous section showed that the automatically produced summaries contained not much less information than those written by humans. In light of this, a normal extension of our method is to automatically translate the summaries and evaluate them in the same manner we evaluated the original summaries. If the automatically translated summaries contain more or less the same information as the Romanian summaries, then the number of correctly answered questions should be the same as in the case of the Romanian summaries.

In order to investigate whether our approach is feasible, the automatically produced summaries were translated from Romanian to English using the free version of eTranslator, an English-Romanian bidirectional translator.[6] Even though the results of the translation were quite disappointing, this program was the only free Romanian to English translation engine we could find on the Internet. All the other programs which we considered either did not work or did not have a free version.

The evaluation of automatically translated summaries was similar to that used to evaluate the Romanian summaries and required judges to answer questions on the basis of summaries they were given. Because we wanted to be able to directly compare the results, the same set of questions as those used in the previous evaluation was used. In order to avoid problems introduced by automatic machine translation, all the questions were manually translated to English. As we did not know how many people we would be able to involve in this experiment, we only evaluated summaries produced by the MMR3 method. The summaries produced by the MMR3 method were chosen because from the point of view of the number of questions correctly answered they are the most similar to human summaries. To our surprise we managed to find 29 judges who answered a total of 414 questions[7]. Table 4 summarises how accurately the judges answered the questions. In order to facilitate comparison, the table also shows the percentage of questions which can be correctly answered on the basis of the Romanian summaries.

Comparison between the two sets of results reveals something which is not unexpected: in all the cases the

[6] The program is available at: http://www.etranslator.ro

[7] In this experiment, some of the judges did not answer all the questions

	Human	Baseline	MMR1	MMR2	MMR3	MMR4
Topic 1	4	2.6	3.5	3.2	3.5	-
Topic 2	3.6	2.5	4	3.1	3.5	3.4
Topic 3	3.9	3	3.4	3.3	3.7	3.8
Topic 4	3.7	3.4	4	3.8	3.9	3.5
Topic 5	3.4	3.1	3.5	3.1	3.2	3.3
All topics	3.72	2.92	3.68	3.3	3.56	3.5

Table 3: The coherence scores assigned by judges to the Romanian summaries

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Total
Accuracy	30%	35%	70%	60%	31%	43%
Accuracy on Romanian summaries	46%	50%	74%	78%	33%	56.53

Table 4: The percentage of questions which were correctly answered by judges who used the automatically translated summaries

percentage of correctly answered questions is lower when automatically translated summaries are used. The overall accuracy of correctly answered questions dropped from 57% to 43%. However, the 43% of questions which can be answered correctly is higher than expected given the poor quality of translations. In fact after reading some of the translations, it is surprising that so many questions could be answered. The percentage of correctly answered questions using the translated summaries is higher than that achieved when the Romanian baseline summaries were used.

Attempts to identify whether a certain category of questions could be answered better then the others failed to reveal any patterns. It does not seem the case that Yes/No questions or questions which required an answer that is not easily deteriorated by the translation process (such as numbers or dates) were easier to answer than the other questions.

Even though on average 43% of the questions could be correctly answered by our judges, their feedback indicates that in most cases without the questions they could not really know what a summary was about. The main reason for this was the poor quality of the machine translation method used which in many cases produced incomprehensible sentences such as the following example: *"Petrovschi stories learned this tax were cooptata in the national lot"*. One of the judges commented that *"The meaning of the texts seemed almost graspable, but just beyond my mental powers."* whilst another compares the texts with a certain character's speech from 'The fast show', a British comedy programme.[8] In light of these comments, it becomes obvious that the translated summaries cannot be used by someone to keep up with what is going on in another part of the world, because even though the texts contain important information, the readers are unlikely to discover as they will give up reading after the first few sentences of the translated text.

No evaluation of coherence was carried out on the translated summaries. The main reason for this was the fact that such an evaluation would have mainly measured the performance of the translation engine.

[8]http://en.wikipedia.org/wiki/The_Fast_Show

6. Conclusions and future directions

This paper has presented a multilingual multi-document summarisation system which can be used to access Romanian news by English speakers. A task-based evaluation where the judges had to answer questions on the basis of summaries shows a decrease in the percentage of correctly answered questions when automatically translated summaries are used. However, after we performed direct as well, we concluded that the decrease is not as large as we expected. This conclusion was reached because some of the translated sentences are barely legible. In this situation, it can be argued that even though the summaries contain the important information, it is unlikely that they would be used by people who do not speak Romanian but want to have access to Romanian news because they are too difficult to read.

In light of this problem, two solutions can be envisaged. The first, and the most obvious one is to use a better machine translation program. An alternative solution, which can also be used together with a better machine translation program, is to extract only sentences which do not have a complicated structure and which can be translated easier. Unfortunately, by employing this method it is likely that important information will be lost, and before such a method is implemented, it is necessary to clearly understand very well how the MT system works.

The advantage of the method investigated in this paper is that it can be easily adapted for any pair of languages as long as it is possible to translate texts from one language to the other. For the future, we plan to experiment with other pairs of languages where there are better machine translation programs such as from English to French or German.

7. Acknowledgements

We would like to thank all the people who kindly accepted to participated in our experiments.

8. References

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. 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, University of Maryland, College Park, Maryland, USA, 20 - 26 June.

Jade Goldstein, Vibhu O. Mittal, Jamie Carbonell, and Mark Kantrowitz. 2000. Multi-Document Summarization by Sentence Extraction. In Udo Hahn, Chin-Yew Lin, Inderjeet Mani, and Dragomir R. Radev, editors, *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, April.

Eduard Hovy. 2003. Text summarisation. In Ruslan Mitkov, editor, *The Oxford Handbook of computational linguistics*, pages 583 – 598. Oxford University Press.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in automatic text summarization*, chapter 23, pages 357 – 379. The MIT Press.

Inderjeet Mani. 2001. *Automatic Summarization*. Natural Language Processing. John Benjamins Publishing Company.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. The MIT Press.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Proceedings of the NAACL/ANLP Workshop on Automatic Summarization*, pages 21 – 29, Seattle, WA, USA, 30 April.

Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001. NewsInEssence: A System for Domain-Independent, Real-Time News Clustering and Multi-Document Summarization. In *Proceedings of the Human Language Technology Conference*, San Diego, CA.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Çelebi, Hong Qi, Elliott Drabek, and Danyu Liu. 2002. Evaluation of Text Summarization in a Cross-lingual Information Retrieval Framework. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, June.

Gerald Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 33(3):193 – 207.

Karen Sparck Jones. 2007. Automatic summarising: The state of the art. *Information Processing and Management*, 43:1449 – 1481.