Multilingual Multidocument Summarization Tools and Evaluation

Horacio Saggion

Department of Computer Science
University of Sheffield
211 Portobello Street - Sheffield, England, UK
saggion@dcs.shef.ac.uk
Tel: +44-114-222-1947
Fax: +44-114-222-1810

Abstract

We describe a number of experiments carried out to address the problem of creating summaries from multiple sources in multiple languages. A centroid-based sentence extraction system has been developed which decides the content of the summary using texts in different languages and uses sentences from English sources alone to create the final output. We describe the evaluation of the system in the recent Multilingual Summarization Evaluation MSE 2005 using the pyramids and ROUGE methods.

1. Introduction

The Translingual Information Detection, Extraction and Summarization (TIDES) program is developing advanced language processing technology to enable English speakers to find and interpret information in multiple languages (`http://www.darpa.mil/ipto/programs/tides`).

With the vast amount of information available in multiple languages, an important problem to be solved is *multilingual text summarization*, the problem of producing summaries in a language T when the input is in a language S different from T or when the input to the summarizer consists of automatic translations in language T of documents in language S. This is a challenging problem because as one can not expect, at least in the near future, perfect automatic translations, the summaries produced from this noisy input would have additional problems to those of lack of cohesion and coherence usually reported in text summarization research (Mani, 2001).

For example, if sentence extracts are produced from automatic translations, then one would find, at the sentence level, grammar and spelling mistakes; additionally if one is to produce "abstracts" instead of extracts from automatic translations, then generic techniques for sentence reduction (Knight and Marcu, 2000) or paraphrase (Barzilay, R. and Lee, L., 2004), which expect good quality input, would have to be adapted.

The 2005 TIDES Multilingual Summarization Evaluation (MSE) concentrates on the problem of generating *multi-document summaries* from a mixture of English and English translations from Arabic documents. The summarization task for MSE 2005 is: given a cluster of topic-related documents in English, create a multi-document summary of the cluster at a given compression rate.

The challenge here consists in extracting content from mixed input in English and automatic English translations of Arabic input.

In order to support experimentation and evaluation, TIDES has created an infrastructure for research groups interested in participating in evaluation, and the Linguistic Data Consortium (LDC) has created clusters of documents in English and Arabic (including translations) which are topically related. In total 50 topic-related clusters have been produced, 25 of them were used for evaluation in MSE 2005.

Topics were obtained from the output of Columbia's NewsBlaster topic clustering system (McKeown et al., 2003) and each cluster was summarized by 4 independent judges at a fixed compression rate of 100 words. The clusters consisting of the English documents, the translations, and the original Arabic documents (sentence aligned to their corresponding translations) were given to participants who had to submit up to three different automatic summaries of no more than 100 words for each cluster.

This paper describes a number of experiments carried out to address the problem of producing summaries from multiple sources in multiple languages. Using the same underlying NLP/summarization systems, we have developed three summarizers which select content from each of the different types of input sources, producing the final summaries relying on one of the sources alone in order to increase readability.

The rest of the paper is organized as follows: In Section 2, we give an overview of our summarization capabilities; Section 3 describes the configurations used for the MSE 2005 evaluation. In Section 4 we present an overview of the pyramid evaluation and the results of the human and automatic evaluation using ROUGE. Section 5 compares our method with previous work and Section 6 closes with conclusions and future work.

2. System

The system used in the experiments described here is an adaptation of a *centroid-based* summarization system. The system for analysis of English documents uses tools for text structure identification, tokenization, sentence boundary detection, named entity recognition, coreference resolution, etc. adapted from the GATE library (Cunningham et al., 2002). The summarization system (Saggion, 2002) implements a number of scoring functions to assess sentence-summary worthiness including sentence position, similarity of the sentence to the document headline, term distribution, named entity distribution, etc. The sentence final score is

This paper is based on a report for the MSE 2005 evaluation.

computed by combining individual feature-values in a linear equation using weights experimentally obtained from training corpora.

Many functions of the system rely on an implementation of the vector space model in which terms are weighted using the formula *term frequency * inverse document frequency* (idf). In general a precompiled idf table (global idfs) is used, where the frequencies may come from a huge text collection; in some cases idfs values computed on-the-fly from the set of documents to be analysed (local idfs) can be used. Vectors of terms are produced for different text fragments including: the whole document, the lead-part of the document (the n% initial tokens of the document, where n is a parameter of the system), and each sentence including the headline (if present). In order to support redundancy detection, the system also computes n-grams for all the input documents.

In a multi-document situation the system receives a cluster of related documents and creates a *centroid* of the cluster in the vector space model (a vector of terms which is in the centre of the document vectors in the cluster) which is used during content selection.

In the case of *generic* multi-document summarization, a unique multi-document summarization feature is combined with two single document summarization features to score sentences.

The multi-document summarization feature is the similarity of the sentence to the centroid of the cluster, measured as the cosine between the two vectors of terms. The single-document features combined with the centroid feature are: (i) sentence *lead-document* similarity, where each sentence receives as value the cosine between the sentence and the lead part of the document (represented as vectors of terms); and (ii) the absolute sentence position, which is a numeric feature with value inversely proportional to the position of the sentence in the document (first sentence is more important, last sentence is less important).

The sentence scores are used to produce a ranked list of sentences from which they are picked up in rank order and included in an extract, unless they are too similar to a sentence already in the extract.

In order to filter out redundant information, the system features an n-gram similarity detection metric. Our approach is shallow in that we use a metric for identifying similar content that relies on n-gram overlap between text units. The metric was used in our DUC 2004 generic multidocument summarization system (Saggion and Gaizauskas, 2004) which had a very good performance in an evaluation of content (2nd system). Having computed n-grams for each document in the input, our n-gram based similarity metric between two text fragments T_1 and T_2 is computed as follows:

$$\sum_{k=1}^{n} w_k * \frac{|grams(T_1,k) \cap grams(T_2,k)|}{|grams(T_1,k) \cup grams(T_2,k)|}$$

where n means that n-grams $1, 2, ... n$ are to be considered, $grams(T, k)$ is the set of k-grams of fragment T, and w_k is the weight associated with the k-gram similarity of two sets. When constructing an extract, a sentence is included in it, only if it is different to all sentences in the extract. In order to implement such procedure, a threshold for our n-gram similarity metric has to be established so that one can decide whether two sentences contain different information. We hypothesise that in a given document all sentences will report different information, therefore we can use the n-gram similarity values between them to help estimate a similarity threshold. We computed pairwise n-gram similarity values between sentences in documents and have estimated a threshold for dissimilar information as the average of the pairwise similarity values.

The adaptation of the system to the Arabic language consisted of the incorporation of: (a) a tokenizer for the Arabic documents, and (b) the creation of inverted document frequencies from the Arabic corpus.

3. System Configurations and Experiments

For the experiments described in this paper, we used three configurations of the system; all of them create summaries by identifying content in different sources and using sentences *from English* documents in order to increase the readability of the output. This was our main concern for the MSE evaluation because we believe that manual evaluation of the content of a summary is affected by its readability. The configurations used work as follows (See Figure 1):

*English summarization (***SYS1***)*: the system takes as input the English documents and extracts sentences from them using the generic summarization system described in Section 2.. The flow of control is shown in Figure 1. The assumption was that because the whole cluster (English + Arabic) would be redundant enough, then one could rely on the English input alone to obtain, at least, some of the main "events" covered in the documents. This assumption was untrue as will be shown in Section 4.3..

*Translated-English summarization (***SYS2***)*: the system takes as input the *translated* Arabic documents and creates an extract by the same method described in Section 2.. As the extract is expected to be of low quality because of its noisy input, then each sentence in the extract (which is an automatic translation from Arabic) is mapped into a true English sentence in the English documents and the result presented as the multi-document extract. Here again, the same assumption about "event" redundancy was made.

*Arabic summarization (***SYS3***)*: the system takes as input the Arabic documents and creates an Arabic extract using the centroid-based summarization system (Section 2.) adapted to work with Arabic documents. Following the method used in (Saggion et al., 2002a), the Arabic extract is first mapped into English translation using the alignment tables, and then the English translation is mapped into an English extract.

Mapping the English translations into the true English sentences is carried out by a process that compares each sentence in the translated extract with each sentence in the English cluster. Each translation in the extract is replaced by the best matching English sentence in the English cluster. Where the best matching sentence is the most similar in

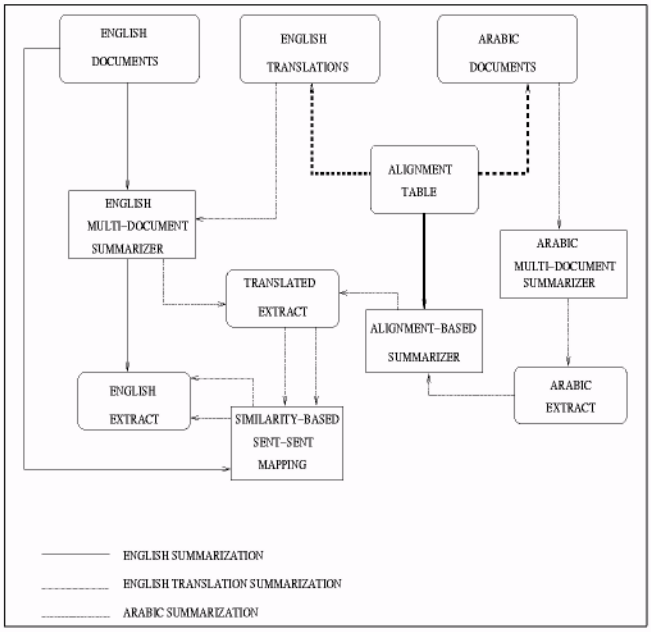

Figure 1: English/Arabic Multidocument Summarization System

terms of our n-gram similarity metric. Here again one has to assume that there is some redundancy of information in the mixed set, and that the main "events" will be present in both the Arabic and English subsets.

4. Evaluation

The summarization method presented here has been evaluated in the recent MSE 2005 evaluation where participants were asked to produce 100-word summaries for each of 25 clusters of topic related documents. A new method for human evaluation of summaries have been recently implemented and tested in MSE. The experiment reported here are a contribution to the first medium-scale evaluation using the pyramid evaluation method. We also provide results of an automatic evaluation using the ROUGE statistic package.

4.1. Pyramid Evaluation

The content of the summaries was evaluated using the Pyramid method (Nenkova and Passonneau, 2004). The method seeks to match content units in peer summaries (e.g., produced automatically) with similar content units found in a pool of human summaries. In this evaluation, a good peer summary is one where its contents units are observed across many human summaries.
In a nutshell, the method is based on:

- (i) the construction of a pool of human summaries for each document cluster;
- (ii) the identification of summarization content units (SCU) in each summary, where content units are proposition-like, atomic representations. For example a text fragment such as "Two Lybians indicted in 1991 for the lockerbie bombing..." will have SCUs such as "two Lybiand indicted", "in 1991", and "bombing";
- (iii) the association of weights to the different SCU based on its frequency of occurrence in the pool of summaries (if a SCU occurs in k summaries, then its weight is k and in an evaluation with s summaries, the maximum possible weight of a SCU is s);
- (iv) the matching of SCU in the human summaries with SCU in peer (e.g. automatic) summaries;
- (v) the calculation of a pyramid formula.

Steps (i)-(iii) give rise to a *pyramid*. A pyramid of order n has n tiers T_i, where each T_i contains SCUs with weight i. Given a pyramid of order n the best possible summary with X units, where content is concerned, is one that contains all SCUs of weight n, all SCUs of weight $n-1$, etc. This is called an *optimal summary*.
In a pyramid of order n, a peer summary will have D_i SCUs appearing in T_i (with $D_i \leq |T_i|$). To evaluate the content of the peer, the following formula is used:

$$\mathtt{D} = \sum_{\mathtt{i}=1}^{\mathtt{n}} \mathtt{i} * \mathtt{D_i}$$

The content value associated with an optimal summary with X SCUs is given by:

$$\mathtt{Max} = \sum_{\mathtt{i}=\mathtt{j}+1}^{\mathtt{n}} \mathtt{i} * |\mathtt{T_i}| + \mathtt{j} * (\mathtt{X} - \sum_{\mathtt{i}=\mathtt{j}+1}^{\mathtt{n}} |\mathtt{T_i}|)$$

where:

$$j = max_i(\sum_{t=i}^{n} |T_t| \geq X)$$

and the final score given to the peer summary is given by the ratio of its score to the maximum possible score $\frac{\mathtt{D}}{\mathtt{Max}}$.
For MSE 2005, ten summaries from ten document clusters (33001 to 33010) were evaluated with the pyramid. We have chosen to evaluate configuration **SYS1**, which had

very good performance in DUC 2004 in a single language task. The average pyramid score of our system is 0.43429 (7th out of 10 systems). In Table 1, we show the scores obtained per cluster and the corresponding system rank. There is great score variability across clusters. In particular, for cluster 33010 our configuration obtained a pyramid of 0 (this was also the case for other 2 systems). If we remove cluster 33010 (outlier) from the evaluation the system's pyramid score goes up to 0.48254 (5th out of 10)- a considerable gain in quality content.

4.2. Automatic Evaluation with ROUGE

Since human evaluation requires human judgements and these are expensive to obtain, automatic evaluation metrics for summary quality have been the focus of research in recent years (Donaway et al., 2000; Saggion et al., 2002a). In particular, the Document Understanding Conferences have adopted ROUGE (Lin.C.-Y., 2004) a statistic for automatic evaluation of summaries. ROUGE allows the computation of recall-based metrics using n-gram matching between a *candidate summary* and a *reference set of summaries*.

ROUGE-n is n-gram recall, ROUGE-L is a recall metric based on the longest common subsequence match and ROUGE-W is a weighted longest common subsequence that takes into account distances when applying the longest common subsequence.

When multiple references are available in an evaluation, the ROUGE statistic is defined as the best score obtained by the summary when compared to each reference. However, the Jackknifing procedure can also be used when M reference summaries are present in the evaluation, this procedure will estimate ROUGE scores by averaging over M sets of $M-1$ possible references. Recent experiments have shown that some ROUGE scores correlate with rankings produced by humans (Lin.C.-Y., 2004). In Table 2 we show the ROUGE scores obtained by our three configurations.

In ROUGE, word overlap is the basis for similarity computation, the problem with this approach is that multi-word units (e.g. "Prime Minister") are not treated as units of meaning while unimportant function words are as rewarding as meaningful content. In order to address this issue, *Basic Elements* (BE) have recently been proposed for automatic evaluation of summaries. These BEs are defined as: the head of a major syntactic constituent; or a relation between a head-BE and a single dependent. So for example for a fragment such as "Two Lybians were indicted", the following BE-Fs will be obtained: `lybians|two|nn` for the entities and `indicted|lybians|obj` for the relation.

BEs can be automatically produced using syntactic analysis together with a set of rules to extract valid BEs from parsing trees. When BEs are extracted using MINIPAR (Lin, 1998), they are called *BE-Fs*. When BEs have been produced for automatic summaries and references, then the BEs of a peer summary can be compared with the BEs of a reference summary to evaluate content in the same way n-grams are used in the conventional ROUGE. Table 3 shows the scores obtained using BE-F as units of meaning.

System	ROUGE	ROUGE (JK)
SYS1	0.05331	0.05352
SYS2	**0.07124**	**0.07129**
SYS3	0.05805	0.05860

Table 3: Rouge scores using BE-F as units of meaning with and without jackknifing procedure.

4.3. Discussion

Where human evaluation (pyramid) is concerned, our configuration **SYS1** obtained an average score. This contrast with the very good performance obtained when our system used all available documents in task 2 in DUC 2004. One possible explanation for this behaviour, still to be verified, is that the sets of documents are unbalanced, with cases where the "main topic" is only described in a fraction of documents. Cluster 33010, for example was problematic for the evaluation, all human summaries concentrate on one specific event described in the cluster "the killing of a palestinian boy" which is described in great length in the Arabic documents but with little detail in the English documents which concentrate on reactions to the event. While the "killing" seems to be a very important event it is unevenly covered in the provided cluster, and as a consequence the system scores very low in cluster 33010.
In general our ROUGE scores are comparable to scores obtained in DUC 2004, however where system ranking is concerned our performance dropped considerably. This is probably the effect of using only English documents to produce the final system output. Our worst configuration according to ROUGE is **SYS1** which produces summaries selecting content from English documents. Surprisely, configuration **SYS3** which selects content from the Arabic documents performs better than the other two when content is evaluated with ROUGE-1 or ROUGE-W, but **SYS2** which selects content from the translations performs better than the other two when the other ROUGE metrics are used including when content is measured using BEs. Interesting, when content is measured using BEs, system **SYS2** obtain reasonable performance which can be attributed to the use of true English sentences to construct the extracts.

5. Related Work

Summarization in languages other than English are not rare (see (Dalianis et al., 2004) for Scandinavian languages and the SUMMARIST project (Hovy and Lin, 1999) for summarization of a variety of languages including Korean and Spanish). In a multi-lingual environment, the first large scale effort for the production of summaries was the focus of a Johns Hopkins Research Workshop (Saggion et al., 2002b) which produced SummBank, the first cross-lingual summarization framework for research in this field. In a restricted domain, the MLIS-MUSI project (Lenci et al., 2002) attack the multilingual summarization problem using symbolic as well as statistical techniques in an information retrieval environment. Statistical techniques including the cue-word, query-sentence similarity, and position methods are used in a process of sentence scoring and selection. Once sentences in the source language have been

Cluster	33001	33002	33003	33004	33005	33006	33007	33008	33009	33010
Pyramid	0.5323	0.6667	0.5278	0.6098	0.3333	0.2258	0.5312	0.5946	0.3214	0
Rank	5	1	3	2	8	10	5	1	3	8/9/10

Table 1: **SYS1** configuration: pyramid scores and system rank (out of 10) per cluster (rank 1 is better).

System	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-W
SYS1	0.33297	0.09476	0.04660	0.03173	0.29388	0.12899
SYS2	0.36740	**0.12636**	**0.07606**	**0.05628**	**0.31489**	0.14083
SYS3	**0.37528**	0.11160	0.05783	0.03915	0.32236	0.14499

Table 2: Rouge scores using n-grams, lcs, and skip bigrams as units of meaning for our three configurations.

selected a linguistic process transforms sentences into semantic representations which are used in a summary generation process in the target language. Open domain multi-document multi-lingual summarization technology is applied in the Newsblaster summarization system (Evans et al., 2004) which used sentence-level similarity computation across languages to cluster sentences, generating the final summary using translated portions of relevant sentences.

6. Conclusions and Future Work

In a multilingual context, summaries are usefull artifacts to overcome the language barrier: cross-lingual summaries produced in the language of the user, for example, can help her assessing the relevance of a document in order to decide whether a good, and sometimes, expensive translation of the source would be required.

The method described here was our first attempt to address the issue of content selection for multi-document summarization in a multilingual environment. Our multilingual prototype was greatly facilitated by the availability of generic NLP tools (Cunningham et al., 2002) and adaptable summarization technology (Saggion, 2002). The results obtained in the evaluation indicate many avenues of further research such as the need to identify content in the whole cluster of provided documents, and investigating possible ways to repair sentences obtained from the translated documents.

7. References

Barzilay, R. and Lee, L. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL 2004*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL 2002*.

H. Dalianis, M. Hassel, K. de Smedt, A. Liseth, T.C. Lech, and J. Wedekind. 2004. Porting and evaluation of automatic summarization. In *Nordisk Sprogteknologi*, pages 107–121.

R.L. Donaway, K.W. Drummey, and L.A. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, pages 69–78. Association for Computational Linguistics, 30 April 2000.

D.K. Evans, J.L. Klavans, and K.R. McKeown. 2004. Columbia Newsblaster: Multilingual News Summarization on the Web. In *Proceedings of NAACL/HLT*.

E. Hovy and C-Y. Lin. 1999. Automated Text Summarization in SUMMARIST. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. The MIT Press.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence*. AAAI, July 30 - August 3.

Alessandro Lenci, Roberto Bartolini, Nicoletta Calzolari, Ana Agua, Stephan Busemann, Emmanuel Cartier, Karine Chevreau, and Jos Coch. 2002. Multilingual Summarization by Integrating Linguistic Resources in the MLIS-MUSI Project. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02), May 29-31*, Las Palmas, Canary Islands, Spain.

D. Lin. 1998. Dependency-based Evaluation of MINIPAR. In *Proceeding of the Workshop on the Evaluation of Parsing Systems*.

Lin.C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization*, Barcelona. ACL.

Inderjeet Mani. 2001. *Automatic Text Summarization*. John Benjamins Publishing Company.

K. McKeown, R. Barzilay, J. Chen, D. Eldon, D. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman. 2003. Columbia's Newsblaster: New Features and Future Directions. In *NAACL-HLT'03 Demo*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of NAACL-HLT 2004*.

Horacio Saggion and Robert Gaizauskas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference 2004*. NIST.

H. Saggion, D. Radev, S. Teufel, and W. Lam. 2002a. Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics. In *Proceedings of*

COLING 2002, pages 849–855, Taipei, taiwan, August 24 - September 1.

H. Saggion, D. Radev, S. Teufel, L. Wai, and S. Strassel. 2002b. Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. In *LREC 2002*, pages 747–754, Las Palmas, Gran Canaria, Spain.

H. Saggion. 2002. Shallow-based Robust Summarization. In *Automatic Summarization: Solutions and Perspectives*, ATALA, December, 14.