# Using SUMMA for Language Independent Summarization at TAC 2011

**Horacio Saggion**
TALN
Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tanger 122 - Barcelona
08018 - Spain
horacio.saggion@upf.edu

## Abstract

The paper describes a language independent multi-document centroid-based summarization system. The system has been evaluated in the 2011 TAC Multilingual Summarization pilot task where summaries were automatically produced for document clusters in Arabic, English, French and Hindi. The system had a reasonable performance in content selection for languages such as Arabic and Hindi and medium performance for English, but poor performance for French. Evaluation results in content selection for French and summary quality in all languages indicate that the system has to be better adapted to the summarization task.

## 1 Introduction

The Text Analysis Conferences (TAC) are a series of evaluation programs to advance the state-of-the-art in various natural language processing problems. Amongst the tasks proposed by the program are knowledge base population from corpora, recognising textual entailment, and automatic text summarization. For our first participation in TAC we have prepared a natural language pocessing system for the Multilingual Summarization pilot task that has as objective to evaluate the application of language independent summarization algorithms. Teams participanting in the task are required to prepare systems able to produce summaries for a number of different languages. More specifically, the task requires to generate a single, fluent, representative summary from a set of documents describing an event sequence.

For the Multilingual Summarization pilot task ten clusters with ten documents each were produced for evaluation, the languages involved in the evaluation were Arabic, Czech, English, French, Greek, Hebrew, and Hindi. We have used our available summarization technology to produce multi-document summaries in four different languages: Arabic, English, French, and Hindi. A single unsupervised system to deal with the four languages was prepared to solve the task. We have made a simplification and assumed that this was a generic text summarization task, as it will be shown and according to the evaluation results this was an over-simplification.

## 2 Multilingual Text Processing

Text processing was carried out using resources available in the GATE system (Maynard et al., 2002). The GATE distribution contains a number of tools for basic text analysis of various natural languages, in particular we have used language specific tokenisers and splitters for Arabic, English, French and Hindi.

## 3 SUMMA Language Independent Summarization Algorithm

For generating the summaries we rely on SUMMA (Saggion, 2008), a set of algoritms for document representation and feature computation implented in Java that uses the GATE Java library. Specific SUMMA components used in TAC (e.g., vector computation, cosine similarity, centroid computation) have already been described in various pa-


Figure 1: Arabic Summary for Cluster 0

pers, so we do not detail them here again. Note that, since there was no data available for adjusting system parameters only default summarization components and some intuitions were used. The following procedure is used to create document representations and score sentences in a cluster of documents:

- Statistics on word frequency at document and sentence level are computed for each document in the cluster to be summarised. Words statistics are nomalized according to their distribution in the document sentences (e.g., a kind of inverted sentence frequency).
- Vector representations are created for each sentence in the document where terms are words and weigths are their normalized frequencies.
- A vector representation is created for the whole document also using words and frequencies.
- The centroid of the cluster is computed as the average of all document vectors.
- For each sentence in each document, the similarity bewteen the sentence vector and the centroid vector is computed using the cosine measure. These similarities are used to rank sentences in descending order of similarity to the centroid of the cluster. The similarity to the centroid is the only summarization feature used by our system.

To select the sentences for the summary, we start from the top ranked sentence adding sentences to the summary content until no more space is available (250 words for this task). Before adding a new sentence to the summary, we verify that the sentence is no reduntant (e.g., there is no similar sentence already in the summary), we compare the new sentence to each sentence already selected and if the similarity is below a threshold we keep the sentence, otherwise we continue with the next sentence on the list. The comparison is carried out with the cosine similarity measure (i.e., we compare the sentence vectors) setting the threshold to 0.90 (i.e., if the similarity is greater then 0.90 we drop the sentence).

The selected sentences are sorted according to the document they belong to. Assuming that the documents are sorted, sentences from document $Di$ are placed before sentences from document $Dj$ if

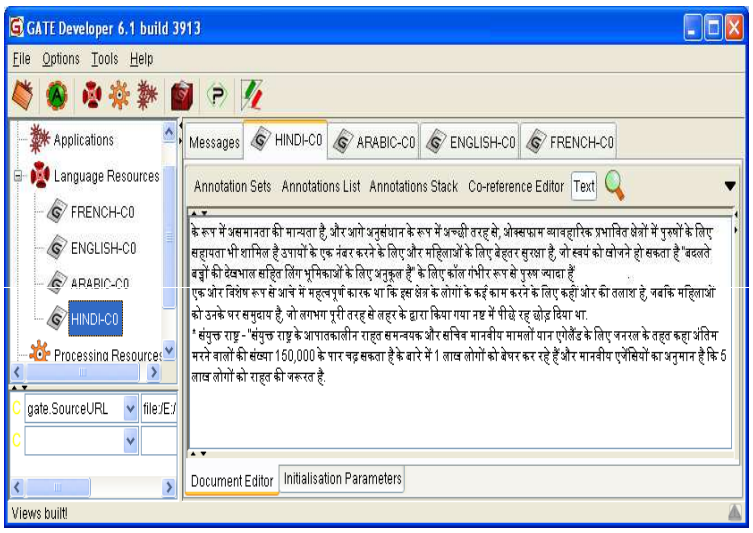


Figure 2: Hindi Summary for Cluster 0

$i \leq j$. If two sentences $Sk$ and $Sl$ come from the same document, then we placed them in document order $Sk$ before $Sl$ if $k < l$ and $Sl$ before $Sl$ otherwise. Documents are sorted according to their document id (Dnnnn).

Examples of summaries created for Arabic and Hindi are shown in Figures 1 and 2 (they are shown in the GATE gui).

## 4 Evaluation Results

System evaluation was carried out using the well-known summarization evaluation package ROUGE (Lin, 2004). It includes a number of measures based on n-grams comparison between a candidate summary and one or several reference summaries. There are several variants of ROUGE in this evaluation two measures were used: **ROUGE-$n$** which is based on a simple comparison of n-grams and **ROUGE-SU4**: which takes into account unigrams in addition to skip bi-grams of a maximum length of 4. For content evaluation, we report results for ROUGE-1, ROUGE-2, and ROUGE-SU4 in Table 1. Three model summaries were available for each document cluster.

| Language | R-1 | R-2 | R-SU4 |
|---|---|---|---|
| Arabic | 0.27 (4/9) | 0.10 (3/9) | 0.12 (4/9) |
| English | 0.39 (7/10) | 0.11 (8/10) | 0.16 (6/10) |
| French | 0.42 (8/9) | 0.11 (9/9) | 0.15 (9/9) |
| Hindi | 0.09 (2/9) | 0.01 (4/9) | 0.01 (5/9) |

Table 1: ROUGE Content Evaluation and Ranks

For the quality of the generated text, human assessment is used to grade each summary with a score of 1 (low) to 5 (high) (three humans assess each summary and all scores are averaged). Results of the evaluation are reported in Table 2 where the length adjusted grade is shown (summaries shorter than 240 words are somehow penalized).

| Language | Len.Adj.Grade |
|---|---|
| Arabic | 2.76 |
| English | 2.20 |
| French | 1.62 |
| Hindi | 1.64 |

Table 2: Summary Quality Evaluation

Ten systems took part on the evaluation (but not all of them produced summaries for the languages we report here): one baseline, one top-line, and eight

peer systems.
Where content of the summaries is concerned our system had a good performance on Arabic and Hindi documents, a medium performance for English, and a poor performance for French.
Were summary quality evaluation is concerned, the grades attributed to our system are too low.

## 5 Outlook

The MultiLing task requires systems to generate a single, fluent, representative summary from a set of documents describing an event sequence. We have treated the summarization problem as generic summarization and prepared a centroid-based summarization system which generally should have an acceptable performance (Radev et al., 2000; Saggion and Gaizauskas, 2004). However, the evaluation results indicate that our system has to be improved in both content and form. This will require better analysis of the input in order to detect not only the events that make up the cluster but also their logical and temporal structure which may help produce a well connected summary.

## References


C.-Y. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization*, pages 74–81, Barcelona, Spain. ACL.

D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. 2002. Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.

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

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

H. Saggion. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Langues*, 49(2):103–125.