FEMsum: A Flexible Eclectic Multitask Summarizer Architecture evaluated in Multidocument Tasks [⋆]

.

Maria Fuentes, Horacio Rodríguez, Jordi Turmo

TALP Research Center, Universitat Politècnica de Catalunya Barcelona, Spain

Abstract

This article describes two types of summarization approaches integrated in a flexible architecture for multitask summarization. The first type is based on the use of lexical features, while the second one is grounded on syntactic and semantic information. All the approaches have been evaluated in experiments where, given a set of documents, they are expected to produce summaries answering a user need (expressed by a query) in a reduced set of relevant textual fragments. Their performance is analyzed in two different tasks: written news and scientific oral presentations.

Keywords: Text Summarization, Spontaneous Speech Summarization.

1 Introduction

Large amounts of digital information are produced on a daily basis in the context of human interaction events, such as news reports, scientific presentations, meetings, etc. Documents generated from these events can be of different nature in terms of their media (e.g., written, audio, video), their domain (e.g., journalism, politics, research, business), or the scenario they originate from (e.g., newspaper, telephone conversation, political speech, business meeting,

[⋆] This research has been partially funded by the European Commission (CHIL project, IST-2004506969) and the Spanish government (TIN-2004-0171-E). We would like to thank all CHIL evaluators. We are especially grateful to Daniel Ferrés, Enrique Alfonseca and Roser Saurí.

Email addresses: mfuentes@lsi.upc.edu (Maria Fuentes), horacio@lsi.upc.edu (Horacio Rodríguez), turmo@lsi.upc.edu (Jordi Turmo).

or scientific conference). In this context, automatic summarization can definitely help deal with the increasing amount of available information.

Automatic document summarization strongly depends not only on the original document features, but also on the user needs (e.g., size of the summary, output media, content related to a query). Current state-of-the-art work in the field reflects such variability. Most of the research is based on English newspaper and newswire data, like the work developed in the context of the Document Understanding Conference[1] (DUC), in both of its two versions: single document (SDS) and multidocument summarization (MDS). Less effort has been devoted to summarizing spontaneous speech, most of the research focusing on broadcast news, typically read aloud from a written text. Current work on oral presentations tends to be based on one single document, the speech transcript, [7], [5], [2], although some other work focus on directly summarizing the speech signal of lectures [8].

Studies such as Shriberg [12] show that oral communication is harder to process than written text. For that reason, we propose using a multidocument summarizer capable of handling documents from different media types. Combining documents from different media can help counteract not only the difficulties in the processing of oral communication, but also those errors introduced by automatic speech recognizers (ASRs). The current word error rate (WER) in speaker-independent ASRs is around 25% on close-talking microphone, and 50% on far-distance. In this article we present FEMsum, a flexible eclectic multitask summarizer which will be evaluated in two different scenarios: written news and scientific oral presentations. FEMsum is a flexible system capable of dealing with different summarization tasks by combining information from documents of different sort, if available, and that takes into account the user needs as well as the particular features of the documents to be summarized.

Next section gives an overview of FEMsum's general architecture. Section 3 describes the components of our tool. We will see that combining them in different ways leads to distinct summarizer approaches, focusing therefore on MDS tasks. In particular, two different approaches have been explored here. A first one, based on lexical information (LEX), and a second one, which uses a richer semantic representation in order to avoid redundancy and improve the cohesion of the resulting summary (SEM). Section 4 discusses the experimental results obtained from evaluating both approaches in different user-oriented, query-focused multidocument summarization tasks. Finally, we conclude in Section 5 with the conclusions from the research.

[1] http:/www-nlpir.nist.gov/projects/duc/

2 Functional Overview of FEMsum's Architecture

The automatic summarization system presented here is a highly modular and parameterizable system able to deal with different information needs. An overview of the system featuring its basic functionalities and global architecture is depicted in Figure 1. The functional requirements are set by means of a parameter set splitted into input and output settings. Input settings concern the characteristics of the documents to be summarized, while output settings apply to the content and presentation of the summary.

Input settings include the following parameters:

- Domain. The summarizer can be domain independent or domain restricted. In the second case, additional knowledge sources can be included, such as: lists of frequent collocations from scientific papers [5]; or a set of gazetteers to help increase the accuracy and obtain finer classes in the named entities classification task restricted to the geographical domain.
- Document structure. The system can take into account information derived from the document structure (position, title, sections, or available tags).
- Language. Currently English and Spanish are supported. The linguistic processing performance depends heavily on this parameter.
- Media. Different media are considered depending on the scenario to be dealt with (video, audio, well written text or any kind of textual document).
- Unit. Both single documents and collections of related documents are used for SDS and MDS respectively.
- Genre. We consider both genre independent and dependent: journalistic or scientific (papers, spontaneous speech, author notes and slides) options.

Output settings include:

- Content. In order to extract the relevant fragments from the input documents, the system can take into account the words from the associated query (in case there is a natural language question or list of keywords), or all the words in the collection (if there is no such query).
- Size. Number of words of the summary.
- Document restriction. Used for filtering out some types of documents, such as those coming from a specific media or genre.
- Output format. The summary can be presented to the user as text, synthesized voice from text, or as an audio/video recorded segment.

3 FEMsum components

In order to achieve the functionalities presented above, the system is organized in three main components (see Figure 1): Relevant Information Detector (RID), Content Extractor (CE), and Summary Composer (SC). In addition, there is a Query Processing component (QP). Not all the components are needed for all the approaches. In fact, in the experiments reported here only RID and SC are always used.

3.1 Relevant Information Detector

The RID module provides a ranked set of relevant Text Units (TU). The definition of TU depends basically on the input media. For instance, in the case of well written text, the TU is the sentence. Two different strategies are used, depending on the existence of a user query (a NL question or a list of keywords). If such a query exists, for each document set the pronoun reference is solved, the text is lemmatized and indexed, and a Passage Retrieval (PR) software (JIRS [6] in the reported experiments) is used to obtain the most relevant TUs. The system retrieves the passages with the highest similarity between the largest n-gram of the query and the one in the passage. RID returns N TUs from passages related to the query. The default value of N is not fixed, but it is the number of TUs from passages selected in some of the executions based on particular user need. If the summary is not query-driven, the system can use all TUs in the document set, or perform a selection and ranking using a tf*idf metric over the whole set.

3.2 Content Extractor

As can be seen in Figure 2, the CE, consists of three components: a Linguistic Processor (LP), a Candidates Similarity Matrix Generator (CSMG), and a Candidates Selector (CS). Input to CE is the set of N TUs provided by RID. All these TUs are processed by LP. Then, CSMG computes the similarities among them, and the most appropiate ones are proposed by CE to be part of the summary.

3.2.1 Linguistic Processor

The LP is illustrated in Figure 3. It consists of a pipeline of general purpose NL processors performing: tokenization, POS tagging, lemmatization, fine grained named entities recognition and classification (NERC), syntactic parsing, semantic labeling (with WordNet synsets, Magnini's domain markers, and EuroWordNet Top Concept Ontology labels), discourse marker annotation, and semantic analysis. Some of these tools are language dependent (English and Spanish), while others are general tools tuned to a specific language. The same tools are used for the linguistic processing of the RID result and for the query (QP) –when needed. The specific tools to be used in each case depend on the type of TU involved and the input language (see [4] for more details).

Tools used for Spanish include:

- FreeLing, which performs tokenization, morphological analysis, POS tagging, lemmatization, and partial parsing.
- ABIONET, a NERC on basic MUC categories (person, location, organization, others).
- EuroWordNet, used to obtain the list of synsets (without attempting Word Sense Disambiguation), a list of hypernyms of each synset up to the top of the taxonomy, and the Top Concept Ontology class.

Tools used for English include:

- TnT, a statistical POS tagger.
- WordNet lemmatizer 2.0.
- ABIONET.
- WordNet
- A modified version of the Collins' parser which performs full parsing and robust detection of verbal predicate arguments.
- Alembic, a NERC with MUC classes, used in order to boost ABIONET performance.

As a result, sentences are enriched with lexical (*sent*) and syntactic (*sint*) language dependent representations. For each sentence, its syntactic constituent structure (including head specification) and the syntactic relations between its constituents (subject, direct and indirect object, modifiers) are obtained. From *sent* and *sint*, a semantic representation of the sentence is produced, the environment (*env*). The information in each of these components is the following:

Sent provides lexical information for each word: form, lemma, POS tag, semantic class of NE, list of WN or EWN synsets and, whenever possible, derivational information.

Sint contains two lists: one recording the syntactic constituent structure (basically nominal, prepositional, and verbal phrases), and the other representing the dependencies between these constituents.

Env is a semantic-network-like representation computed using a process that extracts the semantic units (nodes) and the semantic relations (edges) holding between the different tokens in *sent*. Unit and relation types belong to an ontology of about 100 semantic classes (as person, city, action, magnitude, etc.), and 25 relations between them (mostly binary, as time_of_event, actor_of_action, location_of_event, etc.). Both classes and relations are related by taxonomic links (see [4] for details) allowing for inheritance. Table 1 provides an example of a sentence environment (*env*), and Figure 4 gives the complete representation for another sentence.

3.2.2 Candidates Similarity Matrix Generator

CSMG is in charge of computing the similarity matrix among candidates. For that purpose, it uses the environment of each candidate TU (sentence in the reported experiments). Environments are transformed into labeled directed graph representation, where nodes are assigned to positions in the sentence and labeled with the corresponding token, and edges are assigned to predicates (a dummy node, 0, is used for representing unary predicates). Only unary and binary predicates are used. Figure 5 is the graph representation of the environment in Table 1.

On top of this representation, a rich panoply of lexico-semantic proximity measures between sentences have been built. Each measure combines two components:

- A lexical component which includes the set of common tokens, i.e. those occurring in both sentences. The size of this set and the strength of the compatibility links between its members are used for defining the measure. A flexible way of measuring token-level compatibility has been empirically set, ranging from word-form identity, lemma identity, overlapping of WordNet synsets, approximate string matching between Named Entities etc. For instance, "Romano Prodi" is lexically compatible with "R. Prodi" with a score of 0.5 and with "Prodi" with a score of 0.41. "Italy" and "Italian" are also compatible with score 0.7.
- A semantic component, computed over the subgraphs corresponding to the set of lexically compatible nodes. Four different measures have been defined:
  · Strict overlapping of unary predicates.
  · Strict overlapping of binary predicates.
  · Loose overlapping of unary predicates.
  · Loose overlapping of binary predicates.

The loose versions allow a relaxed matching of predicates by climbing up in the ontology of predicates, e.g. provided that A and B are lexically compatible, *i_en_city(A)* can match *i_en_proper_place(B)*, *location(B)* or *entity(B)*. Obviously, loose overlapping implies a penalty on the score.

Several ways of combining the simple scores have been considered and tested.

Once an appropriate measure has been selected, we can compute the similarity between every sentence pair.

3.2.3 Candidates Selector

In order to select the candidates, three criteria have been taken into account:

- Relevance (with respect to the query or any other element)
- Density and cohesion
- Anti-redundancy

CS proceeds in the following steps:

Let *Sim* be the similarity matrix, *Candidates* a list of candidate TUs, and *Summary* an ordered list of TUs to be included in the summary.

(1) Set *Candidates* to the list provided by RID component.
(2) Set *Summary* to the empty list.
(3) Set *Sim* to the matrix containing the similarity values between members from *Candidates*.
(4) For each candidate in *Candidates*, compute a score that takes into account the initial relevance score and the values in *Sim*. The score used is based on PageRank, as used by Mihalcea and Tarau [10], but without making the distinction between input and output links.
(5) Sort *Candidates* by this score.
(6) Append the most scored candidate (the head of the list) to the *Summary* and remove it from *Candidates*.
(7) In order to prevent overlapping, the $S\%$ TUs most similar (using *Sim*) to the one selected in the previous step are removed as well from *Candidates*. The $R\%$ least scored TUs are also removed from *Candidates*.
(8) If *Candidates* is not empty go to 4.

3.3 Summary Composer

For the summary composition, two different approaches have been explored. The first one is based on lexical information (LEX). The second one uses a richer semantic representation in order to avoid redundancy and to improve the cohesion of the resulting summary (SEM).

The input set of candidate TUs in the LEX approach consists of those TUs previously detected by the RID component as relevant according to the topic. In contrast, in the SEM approach, the input set consists of those TUs extracted by the CE component. Summary TUs are selected by relevance until the desired summary size is achieved. For each selected TU, it is checked whether the previous sentence in the original document is also a candidate. If positive, both are added to the Summary in the order they appear in the original document.

4 Query oriented MDS evaluation

In this section, we describe the experiments we carried out in order to evaluate our system's performance when dealing with different query-oriented MDS tasks. The evaluation is performed on two tasks. On one hand, Section 4.1 reports the experiments and results of FEMsum in the international DUC 2006 evaluation for written news scenario. On the other hand, Section 4.2 analyzes the results obtained in a similar task within the framework of the CHIL[2] project for scientific oral presentation documents.

The tasks in the DUC 2006 and CHIL frameworks differ in the following aspects:

- query (complex vs. list of keywords),
- summary length (250 vs. 100),
- number of input document (25 vs. about 4),
- evaluation test (50 topics vs 20, 10 topics x 2 queries).
- number of manual summary models (4 abstract based vs. 3 extract based).
- domain (journalistic vs scientific),
- input media (well written text vs. raw text comming from: spontaneous speech, and documents related to an oral presentation),
- genre (written journalism vs scientific spontaneous speech, research papers, and author notes or slides from a presentation).

4.1 Experiments in written news scenario

This section describes the DUC 2006 evaluation framework, our approach settings for each kind of summary produced in our DUC 2006 participation, and the results obtained.

4.1.1 Evaluation Framework

For the DUC 2006 evaluation, we were provided with 50 topics which had been selected to be used as test data. Each topic had assigned a cluster of 25 related

[2] http://chil.server.de/servlet/is/101/

textual news documents, as well as a statement describing the information that could be answered using this document cluster. The topic statement could be in the form of a question or set of related questions and can include background information that the assessor has considered would clarify his/her information need. For each topic 4 manual summaries were produced at NIST.

The DUC baseline was a simply system that returned all the leading sentences (up to 250 words) of the most recent document. All 34 DUC 2006 participating systems and the baseline were evaluated at two levels: manually (Linguistic quality and Responsiveness) and automatically (metrics from the package ROUGE, the Recall-Oriented Understudy for Gisting Evaluation [9]). Manual evaluation scored each aspect of a given summary as 1:very poor, 2:poor, 3:acceptable, 4:good, or 5:very good. In addition, our run is one of the 21 participant systems that were also manually evaluated by means of the pyramid method [11].

4.1.2 FEMsum settings in DUC 2006

Our goal in participating at DUC was to evaluate a number of aspects of our system. We therefore submitted three different kinds of automatic summaries in a single run: one lexically based (LEX), and two semantically based (SEM150, SEM250). Out of the 50 summaries we were expected to submit, 7 were produced using the LEX approach, 13 by means of the SEM150 strategy, and 30 by using the SEM250 one. Our system was assigned the identification number 19.

Given the query, a common crucial step in all the approaches is to detect the most relevant TUs (sentences in this experiments). We decided to fix a maximum number of sentences detected as relevant by RID. For that reason we use the corpus of sentences detected as part of a manual summary in DUC 2005 proposed by Copeck and Spakowicz [3]. Analyzing Precision and Recall N was empirically fixed in a maximum of 250.

In the LEX approach, relevant sentences are detected by RID and then SC is applied to obtain the summaries. On the other hand, in both SEM approaches the initial criteria of sentence relevance is that sentences from a same document are considered to have a similar relevance, independently of the RID score. In the SEM250 strategy, all the sentences from the RID output are taken as CE input, whereas in SEM150 the input of CE is the cluster of 150 sentences from the first documents in the set. SEM150 tends to reduce the number of documents whose content is candidate to appear in the summary.

4.1.3 Analysis of the results

The main goal for us to participate in DUC 2006 was to analyze how good our different approaches were in both, detecting the most relevant senteces answering a specific user need, and producing a non-redundant, cohesioned text. For that reason, our result analysis focusses on the scores assigned by the NIST assessors to Content Responsiveness, and Non-redundancy Linguistic Quality aspects.

Table 2 shows the results obtained for each linguistic quality aspect that was manually evaluated: Grammaticality, Non-redundancy, Referential clarity and Focus. For each linguistic aspect every two row detail the score obtained in the subset of summaries produced by each of the three FEMsum approaches (LEX, SEM150, and SEM250), and the mean of the participant systems over this same subset. As can be observed, the SEM approaches have a similar behaviour, both obtaining an acceptable performance (around 3) in all the aspects. Moreover, both of them perform especially well in non-redundancy (around 4). In contrast, LEX obtains only 2,43 in non-redundancy and referential clarity.

Content based responsiveness evaluates the amount of summary information that helps satisfy the information need. First column in Table 3 shows the responsiveness mean score and the distance to the mean participant score obtained by: Humans (4,75); the best system (3,08); FEMsum (2,60); one of the systems evaluated in Section 4.2 +www (2,56) presented in [1]; and the baseline (2,04). In Table 3 can be observed that FEMsum and +www are somewhat above the 2,56 participant mean (0,04 and 0,02 respectively).

To analyze the performance of each approach Table 4 in the first column shows the DUC participant score mean in summarizing the set of document clusters we assigned to each of our approaches. The second column gives the score obtained by our approaches, and the last column shows the distance to the mean. Being among the best participants, SEM150 is above the mean in 0,37, obtaining an acceptable performance in content responsiveness (2,92). Table 5 allows us to better understand the performance of each approach. While in 61,5% of the summaries SEM150 was evaluated as acceptable, good, or very good, LEX and SEM150 were evaluated at least as acceptable in 43% of the summaries. It can be considered that the performance of SEM250 is better than the LEX one because 16,7% of the SEM summaries were scored as good.

The difference between SEM250 and SEM150 can be partly explained by the fact that in the CE component we apply the same S and R factor 15% to prevent overlapping and to remove not relevant candidates for both approaches. That means that SEM250 eliminates a larger number of relevant sentences (15% of 250) than SEM150 (15% of 150). At the same time, it can be observed that reducing the number of candidate documents is not a critical issue.

The second system that was used to evaluate summary content is the pyramid-based one. Under this evaluation, the FEMsum global submission obtained a score of 0,185, only 0,003 points under the mean. 20 clusters were evaluated with this methodology: 3 of them produced by LEX, 5 by SEM150, and 12 by SEM250. We decided that the number of summary samples is not enough to analyze FEMsum performance according to this second perspective.

Table 6 presents the ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) scores obtained by the best ROUGE system (24), FEMsum (19), +www (30), and the baseline (1).Differences between the ROUGE scores obtainded by different systems are small. Scores are low even for the best system. At DUC it was concluded that ROUGE measures do not seem to correlate as well for this task as for other summarization tasks.

4.2 Experiments in the scenario of scientific presentations

In this section, we present the internal evaluation carried out within the CHIL project framework. Five different approaches that were manually and automatically evaluated are briefly described, and results analyzed.

4.2.1 Evaluation Framework

The goal in this framework was to have a summarizer integrated in a smart room, where the summaries to be presented to the user would be the fragment of the image/audio file containing the most relevant information, or synthesizing voice from the text summary. Because of that, it was decided to start by evaluating only summary content aspects. ELDA [3] , in charge of the creation of the CHIL test corpus, selected 10 ISL seminars from the corpus (University of Karlsruhe TH-Interactive System Laboratories). Different sorts of documents were collected for each of these speech events –four on average. For instance, a part of the manual transcription of the seminar, additional documents from scientific publications, conference papers, and presentation slides related to the seminar, if available. All documents were converted into plain text. For each set of documents, two queries were generated according to the seminar topic and the documents content. Then, three human assessors were asked to create extract-based summaries to answer each of the two queries of every document set, with a length of approximately 100 words. These manually created summaries were used as models for applying several

[3] http://www.elda.org/

ROUGE metrics. In particular, ROUGE-n have been studied as simple n-gram co-occurrences between system summaries and human-created models. ROUGE-L and ROUGE-W are used to measure common subsequences shared by two summaries. ROUGE-W introduces a weighting factor of 1.2 to better score contiguous common subsequences. And finally, ROUGE-SUn are used to compute Skip Bigram (ROUGE-Sn) with a maximum skip distance of n.

In addition to the automatic evaluation, a manual content responsiveness evaluation has been carried out at UPC, along the same lines as in DUC. Concretely, 20 assessors were asked to score each automatic and manual summary in terms of summary responsiveness content.

4.2.2 Approaches evaluated in CHIL

The approaches that participated in the CHIL evaluation were:

- PostSeg. We assume that adding textual information helps when summarizing spontaneous speech, for that reason the PostSeg lexical chains based SDS [5] has been used as a baseline. This approach extracts segments of about 30 words only from the transcription and has been adapted to be query-driven by increasing the weight of the lexical chain members that appear in the query.
- LEX (described in Section 3). It only uses lexical information. Segments from all documents are candidates to appear in the summary.
- SEM. Due to the fact that textual transcriptions from spontaneous speech are often ill-formed and they not always follow the written syntactic rules, transcription segments have not been considered as summary candidates. Having a small number of documents to be summarized, we decided to use all the TUs output from the RID module. The number of TUs detected as relevant ranges from 67 to 257 TUs (186,75 in average). The CE module settings are: 15% of S and 10% of R.
- LEXnoT. It is as LEX approach, but without considering the transcription segments as summary candidates.
- +www. In the CHIL corpus, the number of documents to be summarized is smaller than in the DUC one. However, since a lot of scientific information is available online, it would be interesting to evaluate the system presented in [1], which also participates in the DUC 2006 evaluation. This system takes into account background information related to the query from the Internet in order to produce the summaries. As presented before in Section 4.1.3, this system obtains similar results to FEMsum in DUC contest.

4.2.3 Analysis of the results

Table 7 shows the ROUGE metric results when comparing 20 extract-based summaries of a set of documents related to scientific presentations, against three human-created summaries. All the participants perform better than the baseline. Looking at the ROUGE measures, it is difficult to determine whether LEXnoT is better or not than +www.

Looking at the content responsiveness results, in Table 8, we see that LEXnoT obtains the best mean score (2,025), while +www and SEM obtain the same score (1,800). The lower score obtained by a MDS approach is the one obtained by LEX (1,775). That means that better mean performance is obtained when not using the transcription in FEMsum approaches as part of the output. To find out how effective is each approach Table 9 shows the percentage of summaries classified by score. On one hand, although SEM mean is the same as +www (1,8 in Table 8), Table 9 shows that the percentage of summaries considered as 'Acceptable' or 'Good' is higher in SEM (20% + 5%) than in +www (15% + 5%). On the other hand, LEX with a lower mean score (1,775) obtains better results than SEM or +www in terms of percentage score with a 35% of acceptable summaries.

4.3 Contrasting the results obtained in both tasks

Comparing the Responsiveness score (see Tables 3 and 8) obtained from human assessors in DUC (4,75) against the one obtained from CHIL assessors (3,48), it seems that the task proposed in the CHIL framework is harder than the one in DUC. Nevertheless, although the score obtained by the best CHIL system (2,025), is lower than the best DUC system (3,08), the distance from the upper bound (the human assessors) is smaller in the CHIL task (3,48 - 2,035) than in the DUC challenge (4,75 - 3,08). In fact, the score gap between human and system performance is smaller when summarizing scientific presentations than when considering a set of written news as input.

Looking at the performance of the proposed approaches in the two different scenarios, we observe that LEX approaches seem to be robust in terms of responsiveness. LEXnoT scores are 2,025 in CHIL (see Table 9) and 2,29 in DUC (Table 5). Given that LEX mainly use the RID component, this means that the detection of relevant information to appear in the summary achieves similar performance for both tasks, oral scientific presentations and written news. However, SEM approaches obtain better results in DUC (2,53 and 2,92) than in CHIL (1,8). More efforts have to be done in order to set the appropriate parameters of the CE module used for the CHIL scenario.

Comparing the ROUGE scores obtained in both tasks, the approaches obtain better performaces in the CHIL scenario (Table 7) than in the DUC one (Table 6). It seems that the approaches detect words from the models much better in the CHIL scenario than in the DUC one. However, looking at the responsiveness values (see Table 8), it can be assumed that the automatic summary represented by the combination of these words is not good enough.

5 Conclusions

This article presents a flexible system capable of dealing with different summarization tasks by combining information from documents of different sort, if available, and that takes into account the user needs as well as the particular features of the documents to be summarized. Within this framework, we have focused on summarizing written news and textual documents from scientific oral presentations, regarding a query.

Two different types of summarization approaches have been explored in this work. A first one, based on lexical information, and a second one that uses semantics in order to avoid redundancy and improve the cohesion of the resulting summary.

The experiments show two main facts. On one hand, using only lexical information similar performances are achieved in both scenarios: written news and scientific oral presentations. On the other hand, adding semantic information significantly increases the performance when dealing with written news. In contrast, there is room for further improvement when adding semantic information for summarizing documents from different media related to scientific oral presentations.

References

[1] Alfonseca, E., Okumara, M., Guirao, J.M., and Moreno-Sandoval, A., Googling answers' models in question-focused summarisation, in NAACL-HLT Workshop (DUC 2006) New York, United States, 2006

[2] Chatain, P., Whittaker, E., Mrozinski, J., and Furui, S. Class model adaptation for speech summarisation. Proc. NAACL-HLT New York, United States, 2006

[3] Copeck, T. and Spakowicz, S., Leaviring Pyramids. HLT-EMNLP Workshop (DUC 2005) Vancouver, Canada, 2005

[4] Ferrés, D., Kanaan, S., Gonzàlez, E., Ageno, A., Rodríguez, H., Surdeanu, M., and Turmo, J. TALP-QA System at TREC 2004: Structural and Hierarchical

Relaxing of Semantic Constraints. Proc. of the Thirteenth Text Retrieval Conference (TREC 2004). Gaithersburg, Maryland, United States, 2004.

[5] Fuentes, M., Gonzàlez, D., Rodríguez, H., Turmo, J., and Alonso, L., Summarizing Spontaneous Speech Using General Text Properties. Proc. Crossing Barriers in Text Summarization Research Workshop to be held in conjunction with RANLP, Sep 24, 2005, Borovets, Bulgary

[6] Gómez, J.M., Montes-y-Gómez, M., Sanchos, E., Rosso, P., A Passage Retrieval System for Multilingual Question Answering, Proc. TSD, 2005, Plzen, Czech Republic, 2005.

[7] Hirohata, M., Shinnaka, Y., Iwano, K., and Furui S., Sentence extraction-based presentation summarization techniques and evaluation metrics. In Proc. of ICASSP2005, Philadelphia, PA, USA, 2005

[8] Hori, T., Hori, C., and Minami, Y., Speech Summarization using Weighted Finite-State Transducers. Proc. Eurospeech2003, 2003

[9] Lin, C. and Hovy, E., Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proc HLT-NAACL2003 Edmonton, Alberta, Canada, 2003.

[10] Mihalcea, R. and Tarau, P., An Algorithm for Language Independent Single and Multiple Document Summarization, in Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Korea, October 2005

[11] Nenkova, A. and Passonneau, R., Evaluating Content Selection in Summarization: the Pyramid Method, NAACL-HLT 2004.

[12] Shriberg, E., Spontaneous Speech: How People Really Talk and Why Engineers Should Care. Proc. Interspeech, 2005. Libon, Portugal.

Table 1. Sample of environment built from a sentence.

Table 2. DUC FEMsum manually evaluation for linguistic quality scores by approach.

Table 3. DUC FEMsum manual content responsiveness score.

Table 4. DUC Content responsiveness scores by approach.

Table 5. DUC Content responsiveness scores distribution by approach

Table 6. DUC ROUGE measures when considering 4 manual summaries as references.

Table 7. CHIL ROUGE measures when considered 3 manual summaries as references.

Table 8. CHIL responsiveness considering 3 human models when evaluating automatic summaries and 2 when evaluating human summaries.

Table 9. CHIL responsiveness scores distribution by automatic system.

Figure 1. FEMsum Global Architecture

Figure 2. Content Extractor modul

Figure 3. Linguistic Processor constituents

Figure 4. Sample of a sentence analysis and the correspondent environment represeentation

Figure 5. Sample of the graph representation of an Environment

Table 1
Sample of environment built from a sentence

”*Romano Prodi* $_1$ *is* $_2$ *the* $_3$ *prime* $_4$ *minister* $_5$ *of* $_6$ *Italy* $_7$”
i_en_proper_person(1), entity_has_quality(2), entity(5), i_en_country(7), quality(4), which_entity(2,1), which_quality(2,5), mod(5,7), mod(5,4)

Table 2
DUC FEMsum manually evaluation for linguistic quality scores by approach.

	Grammaticality		Non-redundancy		Referential clarity		Focus	
	FEMsum	mean	FEMsum	mean	FEMsum	mean	FEMsum	mean
LEX	3,14	3,45	2,43	4,02	2,43	2,83	3,29	3,73
S150	3,00	3,60	4,15	4,19	3,08	3,09	3,77	3,84
S250	3,33	3,59	4,23	4,27	2,77	3,12	3,20	3,42

Table 3
DUC FEMsum manual content responsiveness score.

System(ID)	Score	Mean Distance
Human(A-J)	4,75	2,19
Best(27)	3,08	1,83
FEMsum(19)	2,60	0,04
+www(30)	2,58	0,02
Baseline(1)	2,04	-0,52
Mean(2-35)	2,56	Stdev 0,28

Table 4
DUC Content responsiveness scores by approach.

	Mean(1-35)	FEMsum	Mean Distance
LEX	2,36	2,29	-0,07
SEM150	2,55	2,92	0,37
SEM250	2,58	2,53	-0,05

Table 5
DUC Content responsiveness scores distribution by approach

	1: Very Poor	2: Poor	3: Acceptable	4: Good	5: Very Good
LEX	14%	43%	43%	0%	0%
SEM150	7,5%	31%	31%	23%	7,5%
SEM250	6,7%	50%	26,7%	16,7%	0%

Table 6
DUC ROUGE measures when considering 4 manual summaries as references.

	Best(24)	FEMsum(19)	+www(30)	Baseline(1)
R-2	0,095	0,076	0,067	0,050
R-SU4	0,155	0,131	0,122	0,098

Table 7
CHIL ROUGE measures when considered 3 manual summaries as references.

	SDS	LEX	LEXnoT	+www	SEM
ROUGE-1	0,293	0,309	0,312	**0,333**	0,323
ROUGE-2	0,060	0,092	**0,102**	0,089	0,073
ROUGE-3	0,029	0,056	**0,064**	0,052	0,032
ROUGE-4	0,019	0,043	**0,050**	0,043	0,021
ROUGE-L	0,256	0,272	0,279	**0,289**	0,280
ROUGE-W1.2	0,089	0,098	0,100	**0,104**	0,098
ROUGE-S1	0,057	0,088	**0,097**	0,087	0,067
ROUGE-S4	0,064	0,089	**0,095**	0,094	0,073
ROUGE-S9	0,069	0,095	0,102	**0,103**	0,083
ROUGE-SU1	0,136	0,162	**0,169**	0,168	0,152
ROUGE-SU4	0,102	0,126	0,132	**0,134**	0,115
ROUGE-SU9	0,090	0,116	0,122	**0,124**	0,105

Table 8
CHIL responsiveness considering 3 human models when evaluating automatic summaries and 2 when evaluating human summaries.

M1	M2	M3	SDS	LEX	LEXnoT	+www	SEM
3,625	3,400	3,375	1,250	1,775	**2,025**	1,800	1,800

Table 9
CHIL responsiveness scores distribution by automatic system.

	M1	M2	M3	SDS	LEX	LEXnoT	+www	SEM
1: Very Poor	0%	0%	0%	70%	40%	15%	30%	35%
2: Poor	10%	5%	10%	25%	25%	50%	50%	40%
3: Acceptable	20%	35%	30%	5%	35%	30%	15%	20%
4: Good	40%	40%	45%	0%	0%	5%	5%	5%
5: Very Good	30%	20%	15%	0%	0%	0%	0%	0%

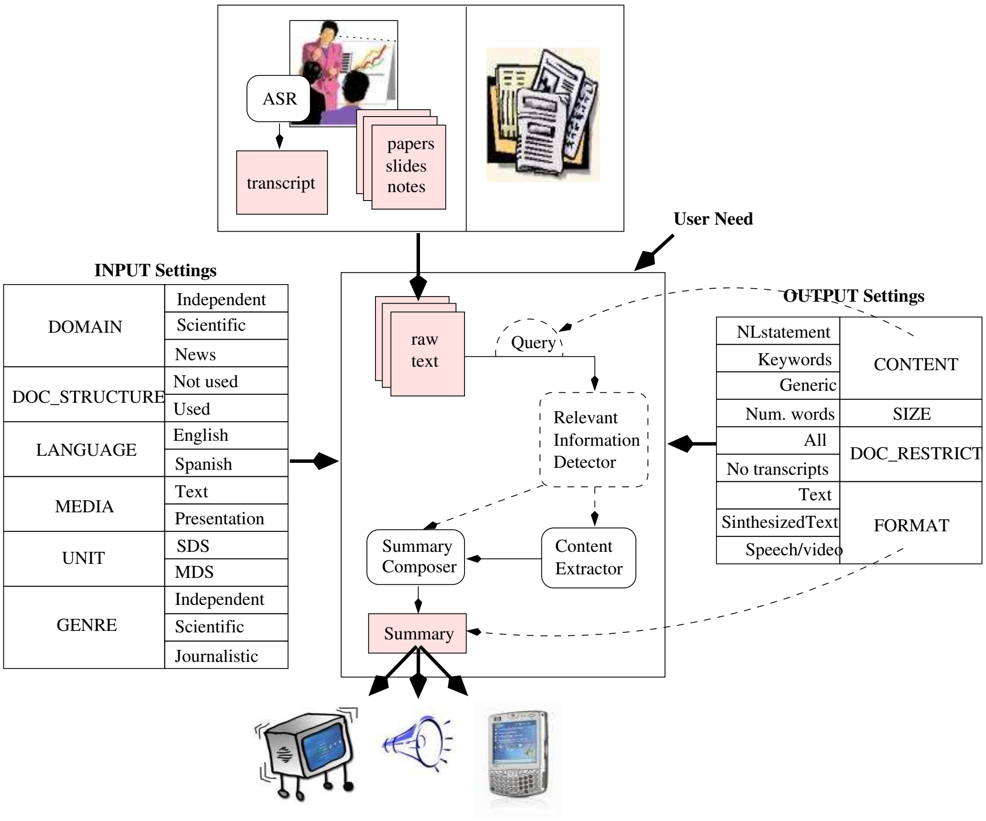

Fig. 1. FEMsum Global Architecture

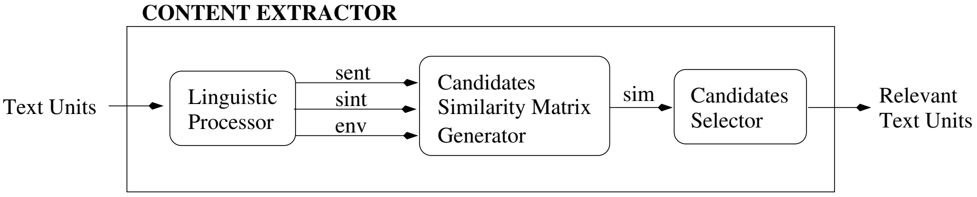

Fig. 2. Content Extractor modul

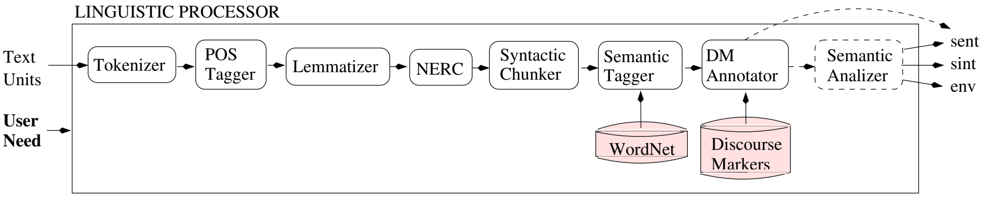

Fig. 3. Linguistic Processor constituents

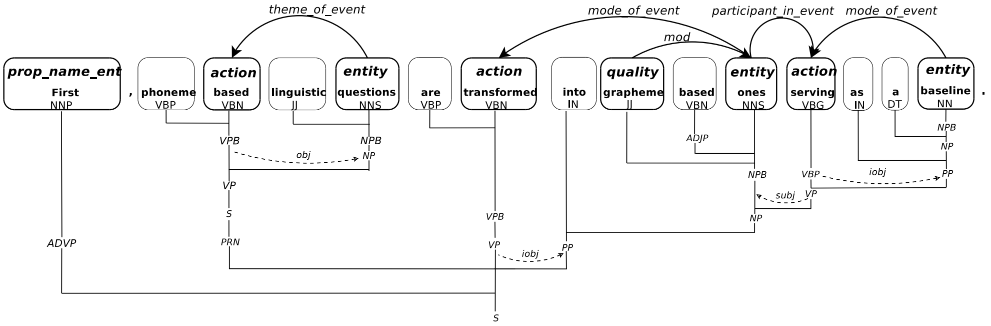

Fig. 4. Sample of a sentence analysis and the correspondent environment represeentation

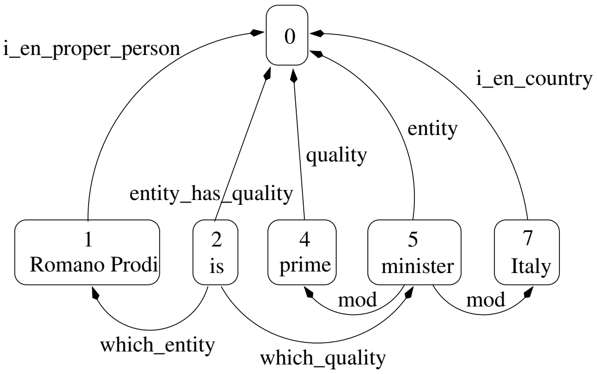

Fig. 5. Sample of the graph representation of an Environment