FEMsum at DUC 2006: Semantic-based approach integrated in a Flexible Eclectic Multitask Summarizer Architecture

Maria Fuentes, Horacio Rodríguez, Jordi Turmo, Daniel Ferrés
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
{*mfuentes, horacio, turmo, dferres*}*@lsi.upc.edu*

Abstract

In order to face different requirements at TALP Research Center we have built a highly parameterized environment allowing to instantiate specific summarizers for different summarization tasks in different languages. This paper describes and analyzes how our system deals with the DUC 2006 task of providing summary-length answers to complex questions. The given query is used to detect relevant passages. After that, semantic similarities between these relevant sentences are detected and then used as input of an iterative graph-based algorithm to avoid redundancy and obtain a cohesioned text. NIST human evaluations are used to analyze several aspects of our system and a specific analysis for each of the three different kinds of submitted summaries is reported.

1 Introduction

In a similar direction as in DUC 2005, DUC 2006 task consists in summarizing a set of documents contributing to answer a user need expressed by several sentences. The main difference with last year contest was that now there is not specification about the summary granularity. The summaries produced by our system are a reduced set of relevant textual fragments extracted from a set of query-relevant passages. As in Question Answering tasks our system uses passage retrieval to detect relevant pieces from the cluster of documents associated to each query. There has been recently a growing interest on applying graph- based representations to NLP tasks, as Question Answering (Molla & Zaanen[6] and Shen et al[9]) or Automatic Summarization (Erkan & Radev[2] and Mihalcea & Tarau[7]). Our approach on facing the summary extraction sub-task follows this line. Instead of using only lexical measures as in [2] and [7], we propose to add semantic measures to establish sentence scores.

In this article we present FEMsum, a flexible eclectic multitask summarizer participating in the DUC 2006 contest. FEMsum is a flexible architecture 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. In the framework of the CHIL[1] project this system is used to summarize different sorts of scientific oral presentation documents.

With the aim of evaluating several aspects of our semantic-based approach we submitted different kinds of automatic summaries in the same run. In our last year participation [4] the Passage Retrieval used did not recover any query-relevant passage for some of the topics. In this case we applied a simple algorithm to produce summaries. We observed that this algorithm obtained better autoPan scores than our system and many other DUC 2005 participants. For that reason, we decided to produce a first kind of summaries based only on lexical features from the passage retrieval (LEX). Two other kinds of summaries (SEM) were produced taking also into account a syntactic and a semantic representation of the sentences and using a graph-representation to establish relations between candidate sentences. The

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

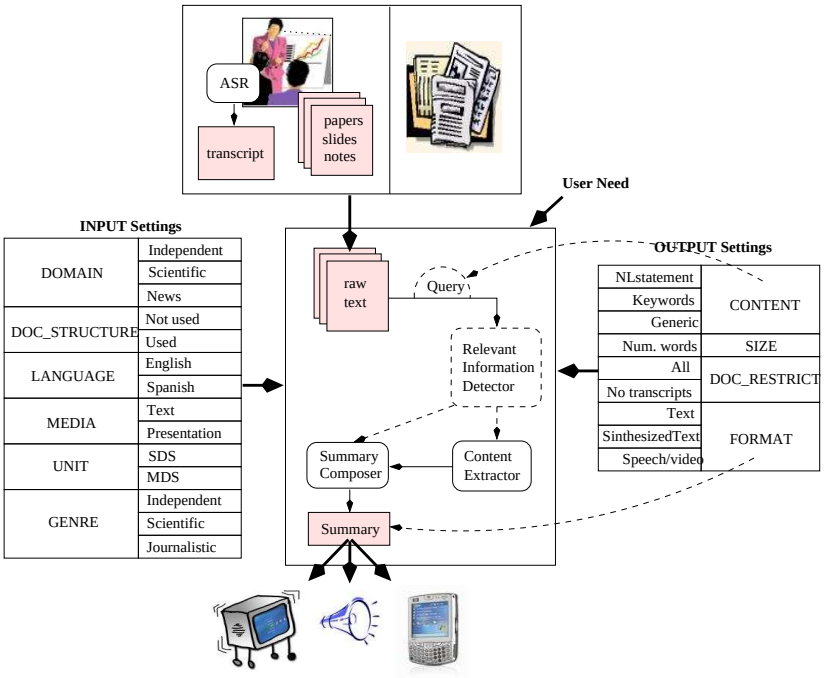

Figure 1: FEMsum Global Architecture

main difference between SEM summaries is the number of sentences considered as summary candidates.

Next section gives an overview of FEMsum's general architecture. Section 3 describes in detail different components of our tool, focusing therefore on DUC 2006 MDS task approaches. Section 4 presents the experimental results and Section 5 the coclusions.

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.

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. 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. In this experiment the TU is the sentence. For each document set the pronoun reference is solved, the text is lemmatized and indexed, and a Passage Retrieval (PR) software (JIRS [5] in the reported experiment) 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.

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 pipe 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 for 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 input language (see [3] for more details).

Tools used for English include: TnT, a statistical POS tagger; WordNet lemmatizer 2.0.; ABIONET; WordNet; and a modified version of the Collins' parser which performs full parsing and robust detection of verbal predicate arguments.

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:

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.

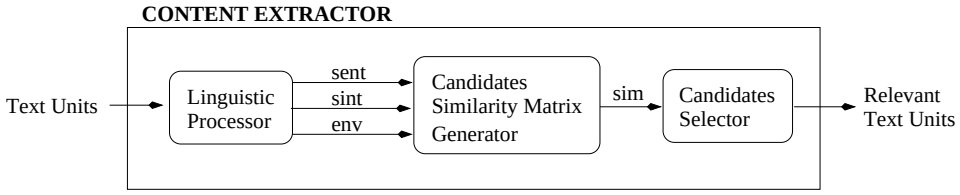

Figure 2: Content Extractor modul

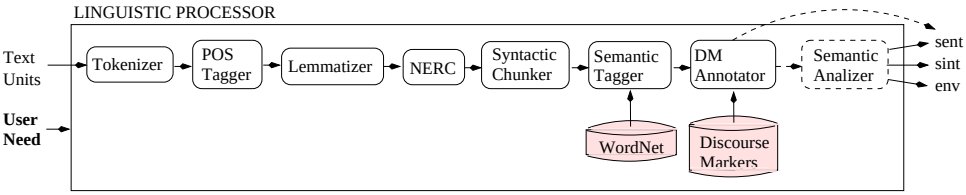

Figure 3: Linguistic Processor constituents

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 [3] for details) allowing for inheritance. Table 1 provides an example of a sentence environment (*env*).

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), quality(4), entity(5), i_en_country(7), which_entity(2,1), which_quality(2,5), mod(5,7), mod(5,4)

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 (sentece 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 4 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 iden-

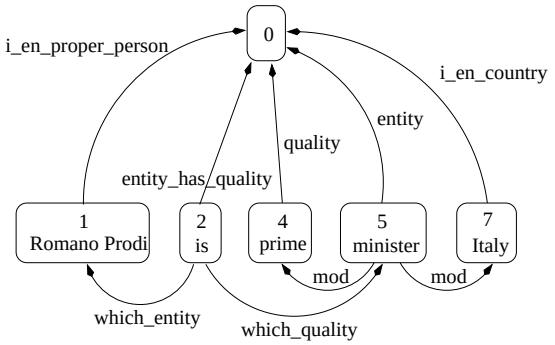

Figure 4: Sample of the graph representation of an environment

tity, 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:

1. Strict overlapping of unary predicates.
2. Strict overlapping of binary predicates.
3. Loose overlapping of unary predicates.
4. 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, *action(A)* can match *human_action(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 criteria), Density and cohesion, and 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 [7], 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 Evaluation

For the DUC 2006 evaluation, we were provided with 50 topics selected to be used as test data. Each topic had assigned a cluster of 25 related 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: Linguistic quality and Responsiveness. 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 [8].

4.1 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 19.

Given the query, a common crucial step in all the approaches is to detect the most relevant TUs (sentences in this experiment). 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 [1]. 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. In the reported experiments any approach processes NERC information.

4.2 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 sentences answering a specific user need, and producing a non-redundant, cohesioned text. For that reason, our result analysis focuses 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 (Q1: Grammaticality, Q2: Non-redundancy, Q3: Referential clarity, Q4: Focus and Q5: Structure & coherence). 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, except in structure & coherence, the aspect with the lowest mean value. Moreover, both of them perform especially well in non-redundancy (around 4). In contrast, LEX obtains only 2,43 in non-redundancy.

Content based responsiveness scores the amount of summary information that helps satisfy the information need. First column in Table 3 shows the responsiveness mean score obtained by: Humans (4,75), the best system (3,08), FEMsum (2,60) and the baseline (2,04). The second column is the distance to the mean participant score (2,56) and the last one the ranking. The global FEMsum submission is some-

Table 2: FEMsum linguistic quality scores by approach, as well as the mean of the 34 participant systems obtained in the associated subset of summaries.

	Q1		Q2		Q3		Q4		Q5	
	FEMsum	mean	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	1,86	2,19
SEM150	3,00	3,60	4,15	4,19	3,08	3,09	3,77	3,84	2,38	2,43
SEM250	3,33	3,59	4,23	4,27	2,77	3,12	3,20	3,42	1,97	2,33

Table 3: Content responsiveness score and mean distance for human, the best system, our submission and the baseline.

System (ID)	Score	Mean Distance	Ranking
Human (A-J)	4,75	2,19	
Best (27)	3,08	1,83	1/35
FEMsum (19)	2,60	0,04	12/35
Baseline (1)	2,04	-0,52	34/35
Mean (2-35)	2,56	Stdev 0,28	

Table 4: 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

what (0,04) above the participant mean.

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 SEM250 summaries were scored as good. Furthermore, the score distribution obtained by the baseline is better than LEX (43%+14%+0% vs. 43%+0%+0%), while the baseline performs worse than SEM250 (13,3%+3,3%+3,3% vs. 26,7%+16,7%+0%).

Table 5: Content responsiveness scores distribution by FEMsum approach contrasted with the baseline.

	1	2	3	4	5
Baseline	43%	0%	43%	14%	0%
LEX	14%	43%	43%	0%	0%
Baseline	23%	38%	23%	8%	8%
SEM150	7,5%	31%	31%	23%	7,5%
Baseline	46,6%	33,3%	13,3%	3,3%	3,3%
SEM250	6,7%	50%	26,7%	16,7%	0%

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 method used to evaluate summary content is the pyramid-based one. As it is shown in Table 6, under this evaluation, the FEMsum global sub-

Table 6: Pyramid evaluation for the best system, our submission and the baseline.

System (ID)	Score	Mean Distance	Ranking
Best (10)	0,257	0,068	1/22
FEMsum (19)	0,185	-0,003	13/22
Baseline (1)	0,121	-0,067	22/22
Mean (1-35)	0,189	Stdev 0,036	

mission 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 the performance of the different FEMsum approaches according to this second perspective.

5 Conclusions

This paper presents a flexible architecture capable of dealing with different summarization tasks. The summary consists in a set of relevant textual fragments extracted from the document set. In our participation in DUC2006 the system is used to answer a user need expressed by a complex question. A passage retrieval software is used to detect the relevant information associated to the user complex question. Lexic-based factual passage retrieval is used to choose the most query related sentences to be summarized. To analyze several aspects we have submitted three different summary approaches. Results show that the use of relevant sentence semantic information helps to avoid redundancy and to obtain better performance in content based measures. One of the approaches ranks among the best participants, obtaining an acceptable performance in content responsiveness and a good performance in non-redundancy.

Acknowledgments

Work supported by the European Commission (CHIL, IST-2004506969) and the Technical University of Catalonia (Daniel Ferrés UPC-Recerca grant). The TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI.

References

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

[2] Erkan, G., and Radev, D., The university of Michigan at duc 2004. NAACL-HLT Workshop (DUC 2004), Boston, MA, United States, 2004.

[3] 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. TREC, 2004, United States, 2004.

[4] Fuentes, M., Gonzàlez, E., Ferrés, D., Rodríguez, H., QASUM-TALP at DUC 2005 Automatically Evaluated with a Pyramid based Metric. HLT-EMNLP Workshop (DUC 2005), Vancouver, Canada, 2005

[5] 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.

[6] Mollá, D., and van Zaanen, M. Learning of Graph Rules for Question Answering Proc. ALTW05, Sydney, Australia, 2005.

[7] Mihalcea, R. and Tarau, P., An Algorithm for Language Independent Single and Multiple Document Summarization, Proc. IJCNLP, 2005, Korea, 2005.

[8] Nenkova, A. and Passonneau, R., Evaluating Content Selection in Summarization: the Pyramid Method, Proc NAACL-HLT 2004, Boston, MA, United States, 2004.

[9] Shen, D., Kruijff, G. J., and Klakow, D. Exploring Syntactic Relation Patterns for Question Answering. Proc. IJCNLP, 2005, Korea, 2005.