FEMsum at DUC 2007

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

Abstract

This paper describes and analyzes how the FEMsum system deals with DUC 2007 tasks of providing summary-length answers to complex questions, both background and just-the-news summaries. We participated in producing background summaries for the main task with the FEMsum approach that obtained better results in our last year participation. The FEMsum semantic based approach was adapted to deal with the update pilot task with the aim of producing just-the-news summaries.

1 Introduction

Automatic Summarization (AS) consists in "to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs" [1]. AS strongly depends not only on the properties of the document, but also on the user needs, such as:

- **Audience.** In case a user profile is available, summaries can be adapted to the needs of specific users, for example, the user's prior knowledge on a determined subject. *Background* summaries assume that the reader's prior knowledge is poor, and so extensive information is supplied, summary teaches about the topic. While *just-the-news* are those kind of summaries conveying only the newest information on an already known subject, assuming the reader is familiar with the topic.
- **Content.** A summary may try to represent all relevant features of a source text or it may focus on some specific ones, which can be determined by queries, subjects, etc. *Generic* summaries provides the author's point of view, considered to be text-driven. While *user-focused* (or query-driven) ones rely on a specification of the user's information need or interest, expressed by a question or a list of keywords.
- **Length.** The targeted length of the summary crucially affects the informativeness of the final result. This length can be determined by a compression rate, that is to say, a ratio of the summary length with respect to the length of the original text. But summary length can also be determined by the physical context where the summary is to be displayed. For example, in the case of delivery of summary news to hand-helds or mobile devices, the size of the screen imposes severe restrictions to the length of the summary.

In DUC 2007, the main task is the same as in DUC 2006: to deal with modeling real-world complex question answering, in which a question cannot be answered by simply stating a name, date, quantity, etc. Given a topic and a set of 25 relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic statement.

The update task consists in producing short (100-word) multi-document update (just-the-news) summaries of newswire articles under the assumption that the user has already read a set of earlier articles. The purpose of each update summary will be to inform the reader of new information about a particular topic. The topics and documents for the update pilot task will be a subset of those for the main DUC task. For each topic, the documents are ordered chronologically and then partitioned into 3 sets, A-C, where the time stamps on all the documents in each set are ordered such that time(A) < time(B) < time(C). There will be approximately 10 documents in Set A, 8 in Set B, and 7 in Set C.

FEMsum summarizer for DUC 2007 is based in our last year best participation approach [2], summaries are produced taking into account a syntactic and a semantic representation of the sentences and using a graph-representation to establish relations between candidate sentences.

Next section presents a description of the FEMsum setting for the main query-focused summarization task and how the system was adapted to deal with the pilot update task. Section 3 presents the experimental results and Section 4 the conclusions.

2 Description of FEMsum and DUC 2007 settings

FEMsum is organized in three language independent components: Relevant Information Detector (RID), Content Extractor (CE) and Summary Composer (SC). In addition there is a language dependent Linguistic Processor (LP) and a Query Processor (QP) component.

As shown in Figure 1, the LP component enriches with linguistic information the original text (documents to be summarized or the user need). This component consists in a pipeline of general purpose Natural Language processors performing: tokenization, part-of-speech tagging, lemmatization, fine grained named entity recognition and classification, syntactic parsing, WordNet based semantic labeling, and semantic role labeling. Textual Units (TU)s, sentences in this experiment, are enriched with lexical (*sent*) and syntactic (*sint*) language dependent representations. For each TU, its syntactic constituent structure (including head specification) and the syntactic relations between its constituents (subject, direct and indirect object, modifiers) are obtained. From

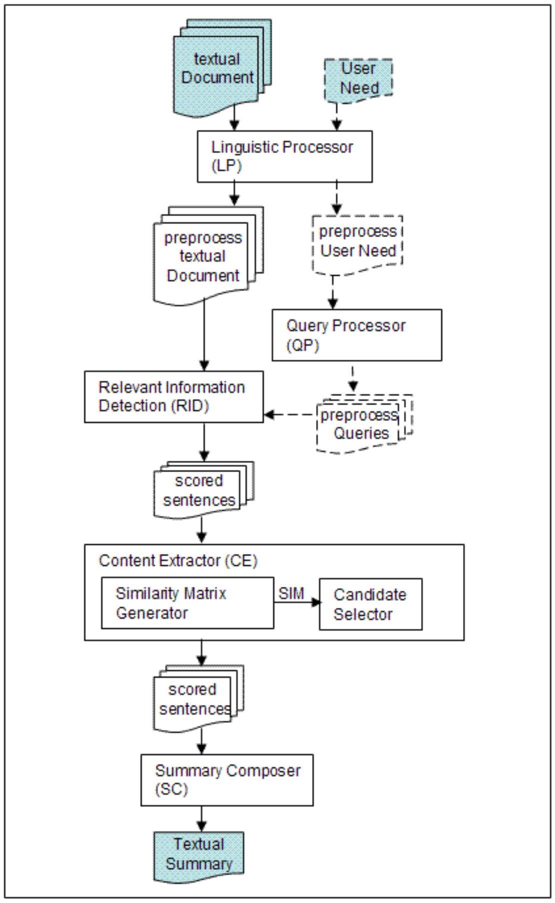

Figure 1: FEMsum data flow

sent and *sint*, a semantic representation of the TU is produced, the environment (*env*).

The RID input is the document or set of documents to be summarized with more or less linguistic information. In the reported experiments, RID instantiation only require stemming. Moreover, the QP output set of queries, expressing the user need, are taken into account by RID to score the set of relevant TUs.

QP involves applying general LP and query specific transformation (as description simplification and query splitting). User's need can be stated in different forms, Figure 2 shows a sample of a natural language complex question and Figure 3 is its QP output.

The linguistic information and the RID output relevance score of the TUs is the input of the CE and SC components. The main function of the CE component is to extract and score by relevance summary candidate TUs.

Section 2.1 describes the new RID component, this is the only last year SEM approach component modified for the main task. Moreover, to deal with update summaries the CE component has also been modified (see Section 2.2.1 for more details). Finally, the SC output is the final text of the summary. In this component is where the summary content post-process is carried out. The post-processing can be more or less elaborate, taking into account the size or the format of the summary. In the SC component the summary TUs can be simplified, paraphrased, reordered or eliminate (see Section 2.3 for more details).

2.1 Relevant Information Detector

As last year, we used the JIRS[1] [3] Passage Retrieval software to obtain the most relevant sentences in the document cluster. In addition, this year we have implemented a TU re-ranking algorithm (see next section).

We used the DUC 2006 corpus [4] to tune empirically the JIRS options as a test set. The Precision, Recall, and F-1 measures were used, giving preference to the options with best F-1 measure.

The following options were tested:

- Retrieval Model. JIRS modes to get passages with a high similarity between the largest n-grams of the question and the ones in the passage are: simple n-gram model, term weight n-gram model, and distance n-gram model. The best retrieval model was the JIRS Distance model with the soft-idf term weighting (distance of 0.1). In this model, the weight of a passage is computed using the larger n-gram structure of the question that can be found in the passage itself and the distances among the different n-grams of the question found in the passage.
- Number of sentences per passage. We experimented with configurations of 1 ,3, and 5 sentences per passage and we obtained the best results with the option of 1 sentence per passage.
- Number of total sentences to retrieve. We tested empirically that the best number of sentences to retrieve was between 100 and 120 sentences.
- Topic fields used to compose the JIRS questions. A retrieval mode that consists in attaching the title at the end of each narrative sentence to compose the queries has achieved better results than the one that uses only narrative sentences alone.
- To filter 'SAY' sentences. Better results were obtained when applying a filter to remove all the sentences that had a form of the verb say after a quoted expression. The other tested filters include: no filtering, filtering sentences with the verb "say", filtering sentences with "say" before a quoted expression, and filtering sentences with say and a quoted expression.

2.1.1 Sentence Ranking

A sentence ranking algorithm has been designed. As a input uses the retrieved sentences for each query from JIRS and a threshold N that indicates the maximum number of final sentences to retrieve. Originally a sentence pool with all the unique sentences retrieved is created. Each sentence in the pool is scored adding the weight of all the passages in which

[1] http://leto.dsic.upv.es:8080/jirs

```
<topic>
<num> D0722E </num>
<title> US missile defense system </title>
<narr>
Discuss plans for a national missile defense system. Include information about
system costs, treaty issues, and technical criticisms. Provide information about
test results of the system.
</narr>
<docs>
XIE19960217.0145
...
</docs>
</topic>
```

Figure 2: An example of the information provided in one DUC 2007 topic.

```
Q1. discuss plan for a national missile defense system ( US ) .
Q2. include information about system cost , treaty issue , and technical criticism
    ( US missile defense ) .
Q3. provide information about test result of the system ( US missile defense ) .
Q4. include information about system cost ( US missile defense ) .
Q5. include information about system treaty issue ( US missile defense ) .
Q6. include information about system technical criticism ( US missile defense ) .
```

Figure 3: DUC 2007 complex natural language query processor output.

it appears. Then, a re-computation of the sentence's weight is applied: if two or more sentences are consecutives in the original document their score is changed with the sum of their weights.

At this point we want to obtain a balanced set of sentences from each query. Then a half of the final N sentences must be selected from the top ranked sentences of each query. If the number of queries is Q, we will obtain the N/(2*Q) top-ranked sentences of each query using as a score the weights computed in the sentence pool.

At last, the unselected sentences of each query are put in a common pool without adding repeated sentences. From this pool the remaining half of N sentences are obtained by selecting the top-ranked ones using the weights computed in the previous sentence pool.

2.2 Content Extraction

The CE requires two modules, the Similarity Matrix Generator (SMG) and the Candidate Selector (CS). This component requires document TUs enriched with semantic information. Similarities among TUs are computed by SMG and used by the CS component.

Next Section details the adaptation carried out in the Candidate Selector component to deal with the DUC 2007 update task.

2.2.1 Candidate Selector

The first change made in this component regarding the one used in DUC-2006 and the main DUC 2007 task is that while in the previous system the score assigned to each candidate sentence was computed using only the semantic similarities ([2]) between the sentences coming from the RID component without taking into account the relevance assigned by JIRS. This year, for the update task we have linearly combined this semantic similarity score with another coming from the JIRS score. The way we have computed this late score is simply considering a linear decay of the scores of each ranked sentence, i.e. the first sentence returned by RID has a score of 1, the score of the following sentences is linearly decreased until reaching 0 for the last ranked sentence. The weight of the combination has been empirically set to 0.9 for the similarity-based score and 0.1 for the JIRS-based one.

In addition, we have faced the update task using the same methods and tools used in the main task with some modifications related to the antiredundancy process.

The process is performed in three iterations: the initial set of sentences (A), the second set (A+B) and the final set (A+B+C) according to DUC-2007 instructions. The first iteration follows the same approach used in the main task. We select, then, a set of sentences (A) to form the first summary. The iterations 2 and 3 follow the same approach. In both cases a previous set of sentences have been produced (and are assumed to be known by the summarizer), A in the second iteration and A+B in the third. After the first and second iterations an additional antiredundancy step is performed for preventing the duplication of information. In this process sentences having a high overlapping with the content of previous summaries are removed. We need, however, to maintain a minimum number of candidate sentences for performing the CE process. For this purpose we have defined two parameters, an absolute threshold, defining the minimum number of sentences to be selected from each set and a relative threshold defining the percentage of sentences provided by RID (the number of sentences that RID considers relevant, according to JIRS re-ranked score is variable) that have to be selected. These parameters have been empirically set to 10 and 0.5. The minimum number of sentences for CE is set to the maximum of these two thresholds. The antiredundancy process, thus, removes the redundant sentences, according to its own threshold, but leaving at least this minimum. The selection process is the same used as in the main task.

2.3 Summary Composition

In our participation in DUC-2006 and in the main task of DUC-2007 (we had no time to include the improvements reported here) the SC component was very simple. The candidate sentences had been previously ranked and the SC had to select in turn the top candidates until reaching the allowed size for the summary. For the update task we have introduced a new component for reordering the already selected sentences in a way of increasing cohesion. We have used the software provided by [6]. This system computes optimal locally coherent discourses, and approaches the discourse ordering problem as an instance of the Travelling Salesman Problem and it solves this known NP-complete problem efficiently in cases similar to ours using a branch-and-cut algorithm based on linear programming. For using this software we had to provide the costs of transitions between units (i.e. the cost assigned in terms of lack of cohesion when a sentence i is followed by a sentence j. We have computed such costs as the inverse of similarities between the corresponding sentences. The system needs too a cost of initial position, i.e. the cost of placing sentence i in the first place of the summary. We have used in this case the inverse of the score assigned to each sentence.

3 Evaluation Results

In this section, we present the results of FEMsum evaluated for both DUC 2007 query-driven multi-document summarization tasks. Section 3.1 details de performance obtained in the main task (FEMsum was assigned the identified 20), and Section 3.2 presents the performance for the update task (FEMsum identified by 49).

3.1 Main task results

Ten NIST assessors wrote summaries for the 45 topics in the DUC 2007 main task. Each topic had 4 human summaries. The human summarizer IDs are A-J.

Two baseline summarizers were included in the evaluation:

Baseline 1 (summarizer ID = 1): return all the leading sentences (up to 250 words).

Baseline 2 (summarizer ID = 2): CLASSY04[5], an automatic summarizer that ignores the topic narrative but that had the highest mean SEE coverage score in Task 2 of DUC 2004, a multi-document summarization task.

NIST received submissions from 30 different participants for the main task. The participants' summarizer IDs are 3-32.

All summaries were truncated to 250 words before being evaluated.

Table 1: FEMsum linguistic quality scores by approach, as well as the mean of the 32 participant systems obtained in the associated subset of summaries.

	FEMsum	Mean
Q1: Grammaticality	2.87	3.54
Q2: Non-redundancy	3.71	3.71
Q3: Referential clarity	3.36	3.20
Q4: Focus	3.40	3.30
Q5: Structure & coherence	2.83	2.42
Mean	3.13	3.24

Table 1 shows the main task 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). The last row presents the mean of FEMsum linguistic quality aspects and the system participant mean.

Content based responsiveness scores the amount of summary information that helps satisfy the information need. First column in Table 2 shows the responsiveness mean score obtained by: Humans (4.71), the best system (3.40), FEMsum (2.93) and the baseline (1.87 and 2.71). The second row is the participant mean score (2.64) and the last column the ranking.

Our system ranks in a 9th position when evaluated with Basic Elements (score of 0.058, 0.008 above the mean). The Baseline 1 ranks in the 29th position and the Baseline 2 in the 19th.

Table 2: Content responsiveness score and mean distance for human, the best system, our submission and the baseline.

System (ID)	Score	Ranking
Human (A-J)	4.71	
Best (4)	3.40	1/32
FEMsum (20)	2,93	8/32
Baseline (1)	1.87	30/32
Baseline (2)	2.71	17/32
Mean (3-32)	2.64	

3.2 Update task results

Ten NIST assessors wrote summaries for the 10 topics in the DUC 2007 update task. The documents for each topic were divided into 3 subsets, A-C, and 4 human summaries were written for each subset. The human summarizer IDs are A-J.

Two baseline summarizers were included in the evaluation:

Baseline 1 (summarizer ID = 35) and Baseline 2 (summarizer ID = 58): CLASSY04 that uses HMM with signature terms as observables and the pivoted QR method for redundancy removal. The sentences are chosen only from the most recent collection of documents. For example, the summary for D0703A-B selects sentences only from the 8 articles in this cluster; however, it uses D0703A-A in the computation of signature terms. Likewise, the summary for D0703A-C selects sentences from only the 7 documents in this cluster and only uses D0703A-A and D0703A-B in the computation of signature terms.

NIST received submissions from 22 different participants for the update task. The participants' summarizer IDs are 36-57.

All summaries were truncated to 100 words before being evaluated.

For the responsiveness evaluation, the assessor for a given topic had previously read all the documents and written a summary for each of the A, B, and C subsets. As a surrogate for rereading all the documents at assessment time, the assessor was given the 4 human summaries for each subset. When evaluating the update summaries for a particular subset of documents, the assessor was reminded that the intended user had already read documents in the earlier subsets. Therefore, information in a summary for subset B that was already in subset A should be discounted; similarly, information in a summary for subset C that was already in subsets A and B should be discounted.

Table 1 shows the results obtained in the update task. Our system performs somewhat under the mean, but obtaining always better results than the Baseline1 (ID=35).

Table 3: FEMsum update task responsiveness evaluation, 22 participants.

	FEMsum	Rank	Mean	40	35	58
A	2.40	16	2.46	3.30	1.80	3.00
B	2.10	17	2.25	2.70	1.90	2.60
C	2.20	12	2.28	2.90	1.30	2.50
all	2.23	15	2.33	2.97	1.67	2.70

4 Conclusions

Last year we produced three different kinds of summary. For the main task, this year we have participated producing the type of summary that obtained the best performance last year. Our system ranks in the top ten for responsiveness producing acceptable summaries.

The main task system was adapted to deal with the DUC 2007 pilot task of generating just-the-news summaries. The main adaptations were carried out when computing redundancy and in the summary composition step were a new reordering algorithm was used. Our system performs somewhat under the mean, but obtaining always better results than Baseline 1. For the update task linguistic quality aspects were not evaluated, for that reason, we are not able to evaluate if applying the new reordering affects in the performance of the system.

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] I. Mani and M. T. Maybury. Advances in automatic text summarisation. MIT Press. 1999

[2] M. Fuentes and H. Rodríguez and J. Turmo and D. Ferrés, FEMsum at DUC 2006: Semantic-based approach integrated in a Flexible Eclectic Multitask Summarizer Architecture. HLT/NAACL Workshop (DUC06), New York, USA, 2006

[3] 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.

[4] T. Copeck and D. Inkpen and A. Kazantseva and A. Kennedy and D. Kipp and V. Natase and S. Spakowicz. Leaviring DUC. HLT/NAACL Workshop (DUC06), New York, USA, 2006

[5] J. M. Conroy and J. D. Schlesinger and J. G. S. and D. P. O'Leary. Left-Brain/Right-Brain Multi-Document Summarization, DUC2004, Boston, USA, 2004

[6] E. Althaus, N. Karamanis, and A. Koller. Computing Locally Coherent Discourses. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)