Chapter 6

Query-focused FEMsum

This chapter presents several FEMsum query-focused Multi-Document Summarization (MDS) approaches evaluated in the most recent DUC tasks. At DUC 2005, 2006 and 2007 contests, both summary content and readability aspects were manually evaluated.

At NIST, the summary content was manually evaluated in terms of summary assessment of responsiveness. Assessors assign an integer between *1* (least responsive) and *5* (most responsive). Responsiveness is different from DUC 2003 and 2004 coverage in that a peer summary is not compared against a single reference. In addition, ISI and Columbia also evaluated automatic summary content taking into account multiple reference summaries. Using the Basic Element (BE) (Hovy et al. 2005) automatic method at ISI and the manual pyramid method (Passonneau et al. 2005) at Columbia.

Apart from summary content, at NIST five linguistic quality scores were assigned to each summary: grammaticality, non-redundancy, referential clarity, focus, and structure and coherence. Each linguistic aspect reflects the degree of a certain readability property to be assessed on a five-point scale from A to E, where A is assigned to a good summary with respect to the quality under consideration and E to a bad one.

Section 6.1 describes each evaluated approach and show how it performs in the complex question-focused DUC summarization main task. In addition, a method to use manually annotated pyramids to automatically evaluate new approaches is detailed and applied to evaluate evolving FEMsum prototypes. In Section 6.2, the best query-focused approach is adapted and evaluated in the DUC 2007 update pilot task, that consists in producing just-the-news summaries. To finish, Section 6.3 concludes the chapter.

6.1 Query-focused Multi-Document Summarization DUC task

The task studied in this section models real-world complex question answering as proposed by (Amigó et al. 2004). The objective is the production of a brief, well-organized, fluent answer to a need of information that could not be met by just stating a name, date, quantity, etc. In 250 words, summaries have to give answer to a question synthesizing a set of related documents.

```
<topic>
<num> d301i </num>
<title> International Organized Crime </title>
<narr>
Identify and describe types of organized crime that crosses borders or
involves more than one country. Name the countries involved. Also
identify the perpetrators involved with each type of crime, including
both individuals and organizations if possible.
</narr>
<granularity> specific </granularity>
</topic>
```

```
<topic>
<num> d426a </num>
<title> Law enforcement with dogs </title>
<narr>
What sorts of law-enforcement tasks are dogs being used for worldwide?
What law enforcement agencies are using dogs? What breeds of dogs are
being used?
</narr>
<granularity> general </granularity>
</topic>
```

Table 6.1: DUC 2005 complex natural language questions.

The task treated in this section was first proposed in DUC 2005. Topics to be used as a test data were created by NIST assessors. For each topic, a set of 25-50 related documents were selected. For each set of documents a title of the topic was established and a narrative statement produced. Table 6.1 presents two DUC topic examples. As can be observed, the topic statement (narrative) is in the form of a set of related open questions. The narrative is a request for information that could be answered using the selected documents. For each topic, it is also indicated the granularity of the answer. The granularity signals if the summary is supposed to name *specific* events, people, places, etc., or in contrast, it is supposed to give a *general* answer.

The most important difference between DUC 2005 and DUC 2006-2007 was that while in the former systems were supposed to produce two different types of summaries, after DUC 2005 only generic summaries were produced. In DUC 2005 it was observed that questions ask for too many diverse subjects (as in the first topic presented in Table 6.1). NIST assessors found that the size limit for the summaries was a much bigger factor in determining what information to include and some *specific* summaries ended up being very *general* (Dang 2005).

FEMsum query-focused MDS prototypes have been manually evaluated in DUC 2005, 2006 and 2007 contests and automatically with the method presented in Section 6.1.1. This method is aimed to study the performance of different approaches taking advantage of DUC judgments.

Section 6.1.2 presents the first experiment carried out with the DUC 2005 data where the first FEMsum prototype, *QAsum*, was evaluated. In Section 6.1.3, the performance of different configurations of *SEMsum* and *LEXsum* participating in DUC 2006 are presented. Then, Section 6.1.4 compares the best scored FEMsum approach evaluated in DUC 2006 with DUC 2005 participants and a new SVM-based approach, *MLsum (MDS, Q).* Section 6.1.5 presents the last *SEMsum* approach evaluated at DUC 2007.

6.1.1 AutoPan, a pyramid-based automatic evaluation method

In DUC 2005, Columbia University constructed pyramids for twenty document clusters (Passonneau et al. 2005), each one with nine summaries written at NIST. Of these, seven were used to create the twenty pyramids and the remaining two were included in the peer evaluation. The authors of the twenty-seven systems evaluated with the pyramid method did the peer annotation. Then, given a pyramid model and an annotated peer the score was automatically computed.

In (Fuentes et al. 2005) we proposed to use DUC 2005 manually annotated pyramids to automatically evaluate new summaries, in a similar direction to (Harnly et al. 2005). To apply AutoPan it is necessary to have a pyramid, as defined by (Nenkova and Passonneau 2004), from each set of documents to be summarized.

As detailed in Section 2.3, the pyramid evaluation method requires a human annotator to match summary text fragments to the Summary Content Unit (SCU)s in the pyramid. AutoPan[1] automates this part of the process as follows:

- The text in the SCU label and all its contributors is stemmed and stop words are removed, obtaining a set of stem vectors for each SCU. The summary text is also stemmed and

[1] Software is available at `http://www.lsi.upc.edu/~egonzalez/autopan.html`

freed from stop words.

- A search for non-overlapping windows of text which can match SCUs is computed. A window and an SCU can match if a fraction higher than a threshold (experimentally set to 0.90) of the stems in the label or some of the contributors of the SCU are present in the window, without regarding order. Each match is scored taking into account the score of the SCU as well as the number of matching stems. The solution which globally maximizes the sum of scores of all matches is found using dynamic programming techniques.

In order to see if the proposed automatic summary constituent annotator correlates with the manual annotations, each one of the 540 summaries manually evaluated with the pyramid method in DUC 2005 is processed with AutoPan. The automatically produced constituent annotations are scored using the same metrics as for manual annotations, and it is found that the scores obtained by automatic annotations are correlated to the ones obtained by manual ones for the same system and summary.

A Spearman test was applyed to the scores obtained by every one of the 540 summaries, including the human ones. The test reports values of $r = 0.52$ for *Original pyramid score* and $r = 0.58$ for *Modified pyramid score*, which exceed the critical value for a confidence of 99%. If we repeat the Spearman test with only the automatic systems (500 examples) the values go down to $r = 0.49$ and $r = 0.52$, but they remain inside the 99% confidence level. Nevertheless, in both cases the scores of automatic annotations tend to be lower than those of manual ones.

We will refer hereforth to the *Original pyramid score* obtained from the automatic pyramids as *AutoPan1* and to the *Modified pyramid score* as *AutoPan2*. The scores obtained from manual pyramids will be referred to as *ManPan1* and *ManPan2*.

If instead of considering the results summary by summary, we take the averages of the scores for each DUC participant system as well as for each one of our approaches described in the following section, there seems to be linear dependency between the two variables. Applying linear regression, we obtain a Pearson coefficient of 0.77 for *AutoPan1* and *ManPan1* and 0.98 for *AutoPan2* and *ManPan2*. Figure 6.1 shows the *ManPan* and *AutoPan* score distribution with respect the obtained regression line.

As summarized in Table 6.2, once found that the *AutoPan* correlated with the *ManPan* ones, it was also analyzed the correlation with the manually assigned responsiveness scores (Resp in Table 6.2), as well as with the ones obtained wiht the ROUGE metrics: ROUGE-2 (R2) and ROUGE-SU4 (RSU4). We apply the Pearson correlation test and the Spearman rank correlation test to the average of the scores obtained in all clusters by every non-human system.

As reflected in Table 6.2, the *AutoPan2* metric correlates well with all other metrics in both tests. All values exceed the confidence level of 99%. *AutoPan1* correlates at 95% confidence level with the manual pyramid-based scores using both tests. The only exception is the 0.360 value

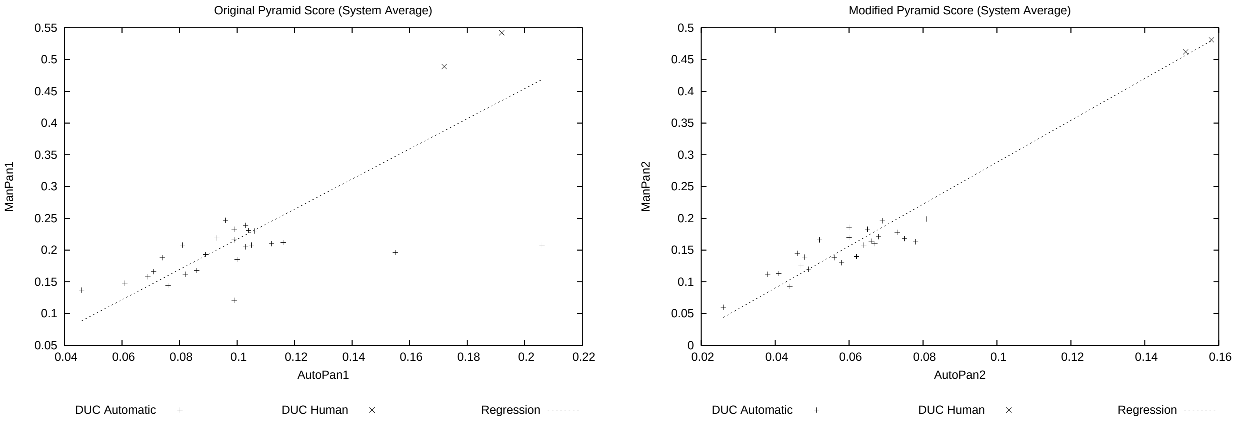


Figure 6.1: Scores obtained by manual and automatic constituent annotation of the human and the participant DUC system summaries, averaged by system.

	Test	R2	RSU4	ManPan1	ManPan2	Resp
AutoPan1	Spearman	0.376	0.330	**0.582**	**0.548**	0.302
	Pearson	0.292	0.238	***0.432***	0.360	0.207
AutoPan2	Spearman	**0.683**	**0.665**	**0.802**	**0.802**	**0.649**
	Pearson	**0.755**	**0.725**	**0.820**	**0.851**	**0.699**

Table 6.2: Correlation values for several metrics, evaluating average scores of non-human systems. Values which exceed the p-value for $p = 0.01$ are shown in **bold**. Values which exceed the p-value for $p = 0.05$ are shown in ***bold italics***.

for the Pearson correlation with *ManPan2*, which does not exceed the imposed 0.05 p-value. For ROUGE measures and responsiveness, no correlation test achieves this confidence level.

One of the explanations to this non-correlation can be seen in Figure 6.2, which describes the behavior of the four automatic measures when evaluating non-human systems. It can be seen that two systems are particularly wrongly scored by *AutoPan1*. These systems, 11 and 26, score at the level of human summarizers. In fact, system 11 is better scored than two humans. It seems that *AutoPan1* must have some sensitivity to the kind of automatic summary produced. However, the plots for *AutoPan2* and the ROUGE metrics have a similar shape, and this is in accordance with the correlation values seen in Table 6.2.

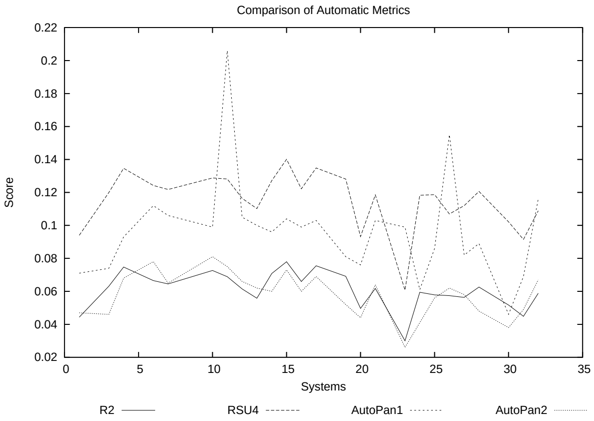


Figure 6.2: Comparison of automatic metrics applied to DUC 2005 participants.

6.1.2 Using the TALP-QA system at DUC 2005

In this section the performance of several *QAsum* prototypes is compared using the DUC 2005 data. The first query-driven MDS FEMsum approach uses extraction techniques to produce summaries, selecting and putting together several portions from the original set of documents. To face the DUC 2005 task, *QAsum* summaries are produced using part of the TALP-QA Question & Answering system.

The TALP-QA is in continuous evolution trying to adapt itself to the specific characteristics of TREC and CLEF contents. When *QAsum* was designed, the last TALP-QA English prototype had participated in TREC 2005 QA track (Ferrés et al. 2005) and the last Spanish prototype in CLEF (Ferrés et al. 2005).

For the experiments reported in this section, TALP-QA system has been used without the specific Definitional QA subsystem. Moreover, as we are not interested in obtaining the factual answer of the questions but only to identify the sentences where it is located, only the initial steps of TALP-QA are performed (Collection and Question Processing, Passage Retrieval and Answer Extractor).

As depicted in Figure 6.3, *QAsum* instantiates five FEMsum components:

1. **English TALP-QA LP**

   Each set of documents is pre-processed with the general purpose linguistic tools described in Section 3.4. Text is enriched in order to obtain the TALP-QA lexico-morphologic, syntactic and semantic representations also presented in Section 3.4.

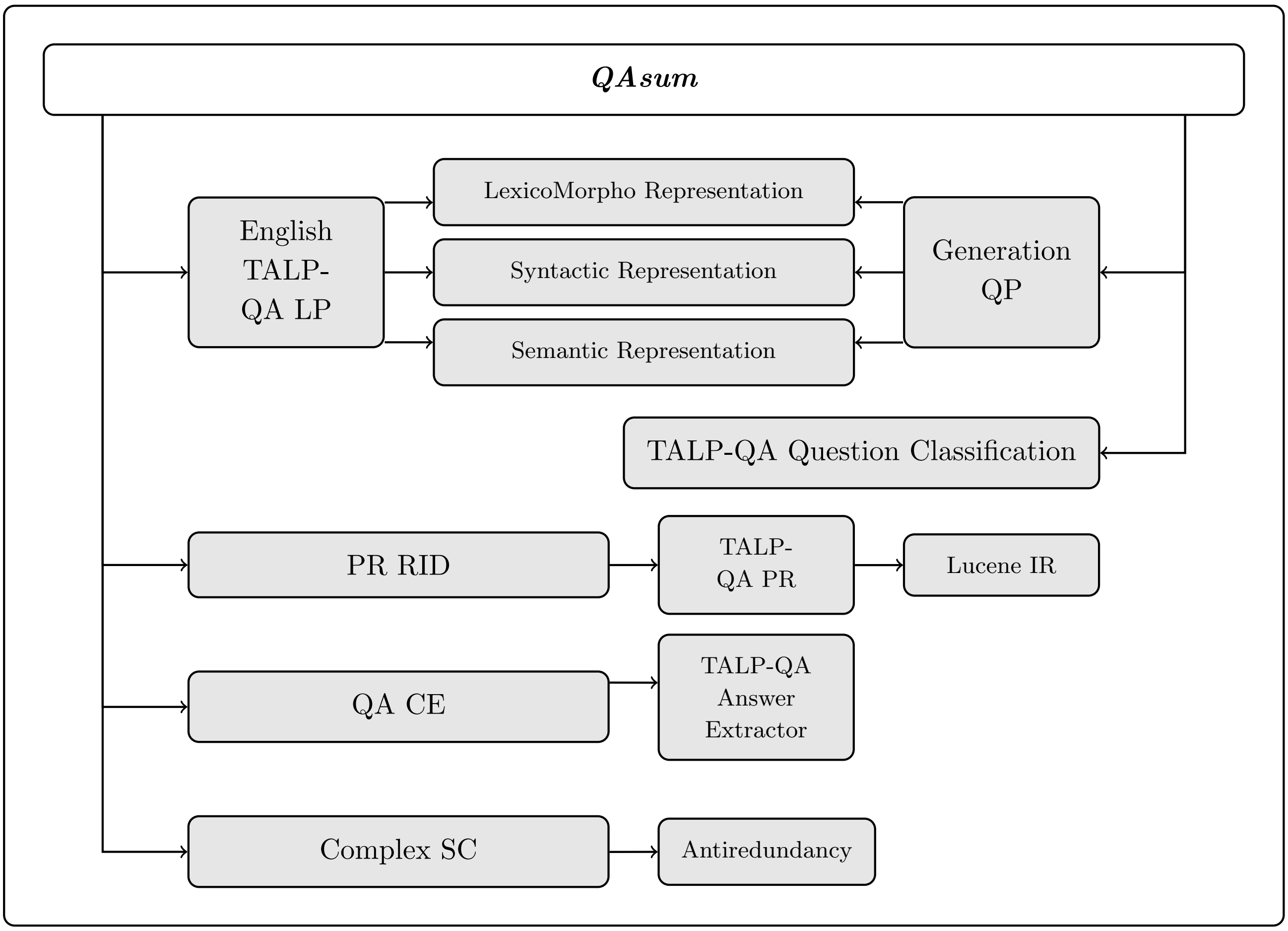


Figure 6.3: Modules of *QASUM-TALP*, a *QAsum* approach.

The enriched collection is indexed using *Lucene*[2] IR engine. Lucene is used to create an index with two fields per document:

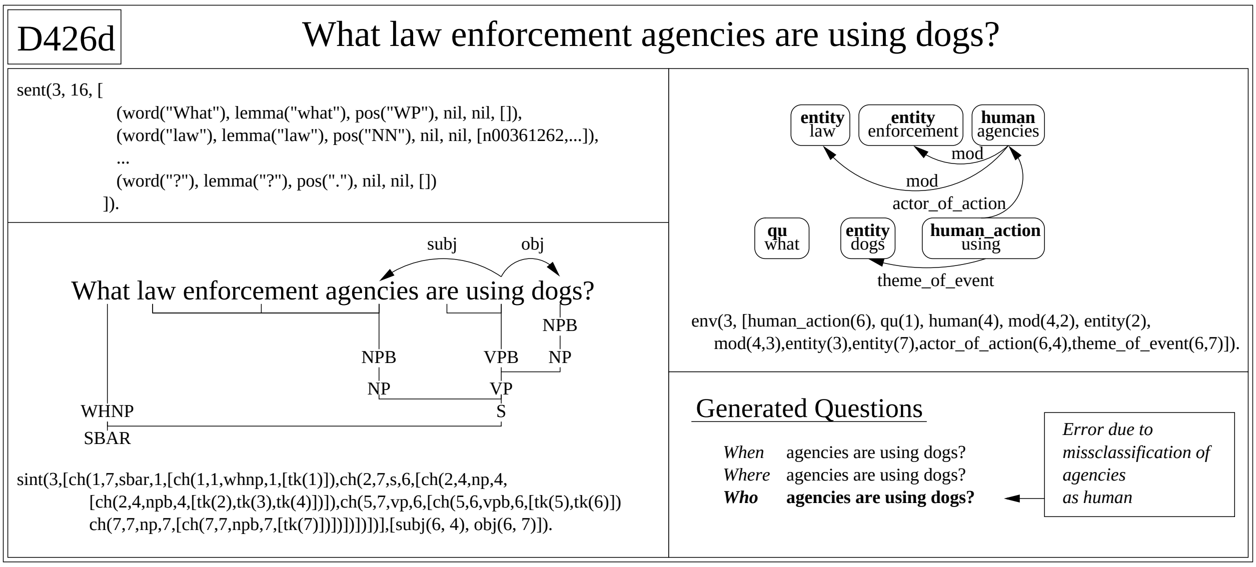
- lemmatized text with NEs recognized and classified and syntactic information
- original text (forms) with NEs recognized (not classified) and syntactic information.

As an additional knowledge an IDF weight is computed at document level for the whole collection. This information is used by the TALP-QA components.

[2] http://jakarta.apache.org/lucene

2. **Generation QP**

   As the current TALP-QA system does not process complex questions, a set of simpler factual questions is automatically generated from the complex description given by the user (the narrative in Table 6.1).



```
sent(3, 16, [
        (word("What"), lemma("what"), pos("WP"), nil, nil, []),
        (word("law"), lemma("law"), pos("NN"), nil, nil, [n00361262,...]),
        ...
        (word("?"), lemma("?"), pos("."), nil, nil, [])
      ]).
```

```
sint(3,[ch(1,7,sbar,1,[ch(1,1,whnp,1,[tk(1)]),ch(2,7,s,6,[ch(2,4,np,4,
      [ch(2,4,npb,4,[tk(2),tk(3),tk(4)])]),ch(5,7,vp,6,[ch(5,6,vpb,6,[tk(5),tk(6)])
      ch(7,7,np,7,[ch(7,7,npb,7,[tk(7)])])])])])],[subj(6, 4), obj(6, 7)]).
```

```
env(3, [human_action(6), qu(1), human(4), mod(4,2), entity(2),
   mod(4,3),entity(3),entity(7),actor_of_action(6,4),theme_of_event(6,7)]).
```

Generated Questions

When agencies are using dogs?
Where agencies are using dogs?
Who **agencies are using dogs?**

Error due to missclassification of agencies as human

Figure 6.4: Example of a sentence linguistically processed and the generated questions.

Figure 6.4 is an example of the TALP-QA lexico-morphologic (sent), syntactic (sint) and semantic (env) representations from one of the sentences in the narrative, *What law enforcement agencies are using dogs?*. The figure also presents the generated questions: *When agencies are using dogs?*, *Where agencies are using dogs?* and *Who agencies are using dogs?*. In this case *using* has been detected as a **human_action**. For that reason, the new questions are generated considering the chunk that contains both the subject *agencies* (**human**) and the object of the action *dogs*. However, that is an example where the utility of the generated questions is uncertain. More details about the Generation QP component are given in Section 3.5.2.

3. **TALP-QA Passage Retrieval RID**

   The main function of the TALP-QA Passage Retrieval (PR) component is to extract small text passages that are likely to contain the correct answer. Document retrieval is performed using the *Lucene* IR system. Each keyword is assigned a score using a series of heuristics. For example, a proper noun is assigned a score higher than a common noun, the question focus word (e.g. *agencies* in the question *When agencies are using dogs?*) is assigned the lowest score, and stop words are removed. The PR algorithm uses a data-driven query relaxation technique: if too few passages are retrieved, the query is relaxed first by increasing the accepted keyword proximity and then by discarding the keywords with the lowest score. The contrary happens when too many passages are extracted.

4. **TALP-QA CE**

   The TALP-QA Answer Extractor consists of two tasks performed in sequence: Candidate Extraction and Answer Selection. However, for summarization purpose only the first component is used to extract all the candidate answers from the highest scoring sentences of the selected passages. As described in Section 3.7.3, sentence are extracted taking into account the question type previously extracted from the question text.

5. **Complex SC (Antiredundant)**

   The sentences included in the final summary are selected one at a time from the set of sentences retrieved in the previous phase. As described in Section 3.8.1, taking into account the semantic representation, the redundancy of the candidate sentence is measured with respect to the previously selected ones. When all the retrieved sentences have been considered and the size has not been reached, the first sentence of each document crossed by a lexical chain is considered until achieving the desired size.

Analysis of the results

A first *QAsum* approach, *QASUM-UPC*, was manually evaluated at DUC 2005 (Fuentes et al. 2005). We detected that in the LP step NEs were not used in the index creation. The other detected malfunction in the submitted approach was that NEs from complex question were sometimes badly segmented due to the few context for recognition. Both problems were solved and a second set of summaries, *QASUM-TALP*, were computed.

Using the black box evaluation presented in Section 6.1.1, the new approach, *QASUM-TALP*, was compared with two baselines, *DefQA*, the TALP-QA Definitional QA subsystem (Ferrés et al. 2005) and *LC*, the SDS FEMsum approach preseted in Chapter 4. *LC* is used in the Antiredundant SC when there are not enough retrieved sentences.

As for the new approaches *ManPan1* and *ManPan2* were not available, we used the obtained regression line to hypothesize what their manual scores would be from the automatic scores *AutoPan1* (left plot in Figure 6.5) and *AutoPan2* (right plot in Figure 6.5). These plots were obtained taking into account the averages of the scores for each DUC participant system as well as for each one of our evaluated approaches.

As said in the previous section applying linear regression, the obtained Pearson coefficient for AutoPan1 and ManPan1 is 0.77 and for AutoPan2 and ManPan2 is 0.98.

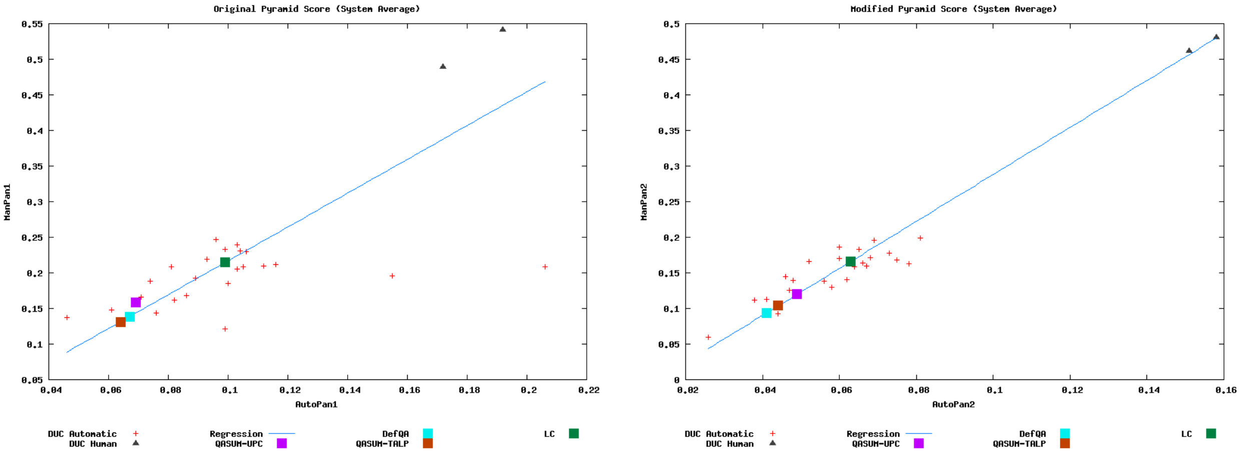


Figure 6.5: AutoPan scores of the DUC systems and the new approaches, averaged by system.

As depicted in Figure 6.5, the best results are obtained by *LC*, which reaches an average position with respect to the rest of DUC systems. *DefQA* obtains slightly lower results than *QASUM-UPC*, the submitted approach, and the second *QAsum* approach, *QASUM-TALP*, scores worse than the submitted approach. As *QASUM-TALP* was still a preliminary version, there can be multiple reasons for this surprising decrease in the performance.

In order to explain why the first version performs better than the second one a glass box evaluation was carried out. First of all it was observed that 23 of the 50 submitted summaries were produced using only the lexical chain information, *LC*, because no candidate sentence was selected by the TALP-QA CE component. As depicted in Figure 6.5, *LC* summaries are quite well scored. Taking into account the subset of 20 clusters with manual Pyramid information, we compare the number of *LC* summaries produced by *QASUM-UPC* with the number of *LC* summaries produced by *QASUM-TALP*. 8 of the 20 summaries produced by *QASUM-UPC* were *LC*, using only the lexical chain information, while only 1 *QASUM-TALP* summary was of this type.

To check how well did the TALP-QA system extracting relevant information, the summary sentence corpus proposed by (Copeck and Szpakowicz 2005) was used. This corpus takes into account the information from the pyramid manual evaluation method in order to determine which document sentences are candidate to appear in a summary. The corpus was aligned at our sentence segmentation level to check how good was the TALP-QA PR module extracting the relevant sentences. Then, we study the Precission (P) and the Recall (R). Table 6.3 shows the results obtained.

Evaluated with data from TREC'05, the TALP-QA PR obtains an accuracy of 62,60% (216/345) of questions for which an answer has been found within their set of passages (Ferrés et al. 2005). Comparing this accuracy with the F-measure presented in Table 6.3 (0.27 for

Approach	P	R	F
QASUM-UPC	0.27	0.28	0.27
QASUM-TALP	0.37	0.24	0.29

Table 6.3: Precision, Recall and F-measure obtained when selecting relevant sentences.

QASUM-UPC and 0.29 for *QASUM-TALP*), it can be considered that the performance of TALP-QA PR decreases when working in the DUC 2005 task. It can be concluded that TALP-QA is affected by the fact that it was designed to answer factual questions as defined in TREC. Checking the generated questions it was observed that often they were not useful. That fact is probably reflected in the decrease of TALP-QA PR performance.

6.1.3 Comparing several FEMsum approaches at DUC 2006

In the first experiment reported in the previous section, the PR used did not recover any query-relevant passage for some of the topics. Moreover, applying a simple algorithm to produce summaries obtained better AutoPan scores than our approaches and many other DUC 2005 participants. For that reason, we decided to produce a first simple type of summary based only on lexical features, *LEXsum*. This first approach is simpler than *SEMsum* that follows the growing interest on applying graph-based representations to NLP tasks. However, instead of using only lexical measures as (Erkan and Radev 2004) and (Mihalcea and Tarau 2005), *SEMsum* takes into account semantic measures to establish sentence scores.

Our goal in participating at DUC 2006 (Fuentes et al. 006a) was to evaluate a number of FEMsum aspects. 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.

Following the FEMsum architecture depicted in Figure 6.6 it can be observed that *SEMsum* approaches are organized in five main components. As reflected in the figure in grey, *SEMsum* and *LEXsum* share three components. The most important difference is that *LEXsum* does not instantiate the SEM CE component. For that reason, while, as *QAsum*, *SEMsum* instantiates the TALP-QA LP, a simpler LP component is instantiated by *LEXsum*.

1. **English TALP-QA LP**: While for *SEMsum* approaches all the linguistic modules described in Section 3.4 are used to obtain the TALP-QA lexico-morphologic, syntactic and semantic representations, in *LEXsum* only text stemming is needed. For each document set the pronoun reference is solved, the text is lemmatized and indexed. Any approch evaluated in DUC 2006 contest used the NERC module.

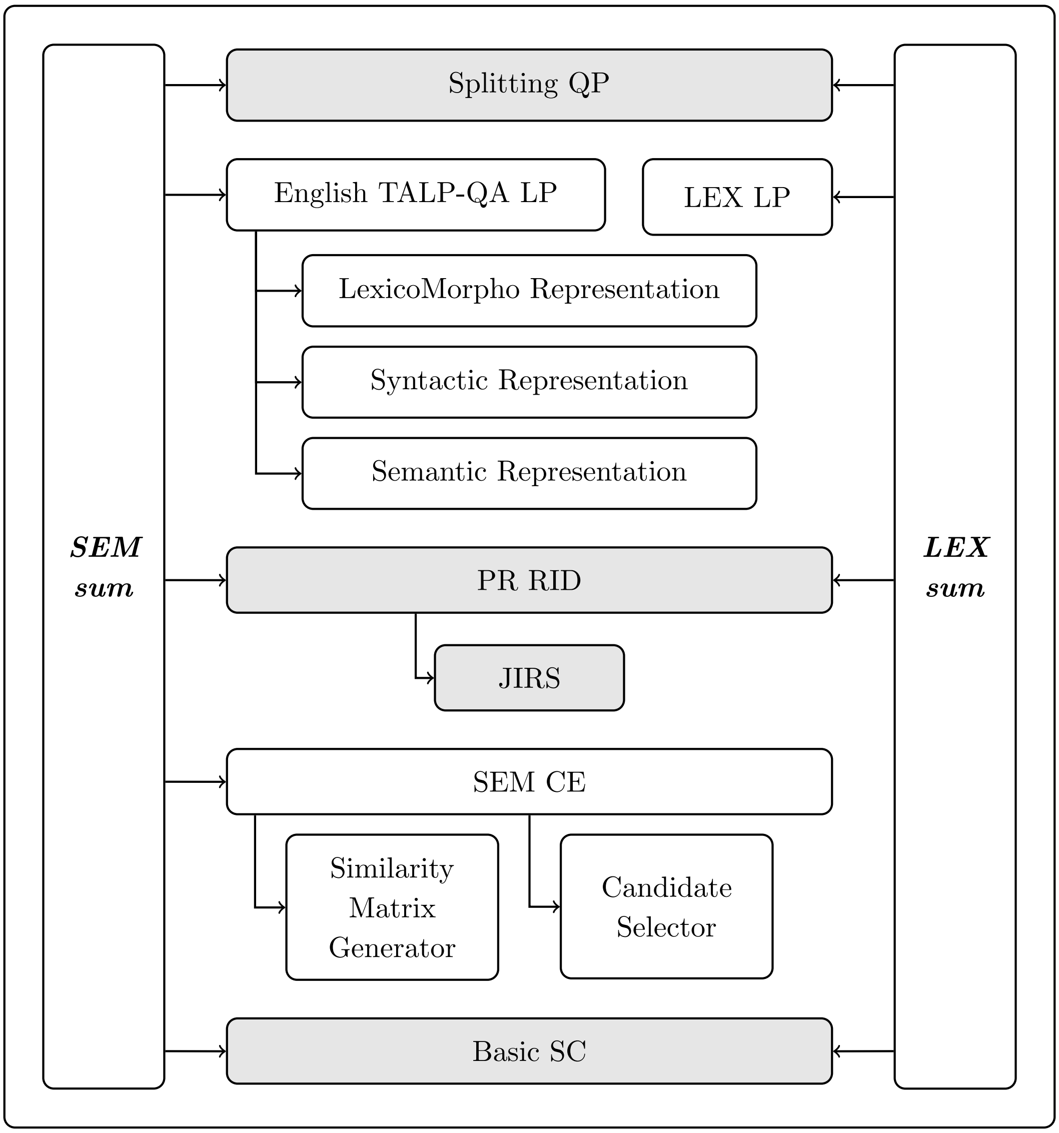


Figure 6.6: FEMsum approaches at DUC 2006.

2. **Splitting QP**

   Each sentence from the narrative is considered as a question to be taken into account, as well as the ones produced by the splitting process described in Section 3.5.1. All words of the title not present in the original sentence are added at the end of the respective question.

3. **JIRS Passage Retrieval RID**

   The PR RID instantiated is described in Section 3.6.3 and uses the JIRS PR software to obtain the most relevant TUs. In the experiments reported in this section the TU is the sentence. 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 sentences from passages related to the query. We use the corpus of sentences detected as part of a manual summary in DUC 2005 proposed by (Copeck and Szpakowicz 2005) to analyze JIRS Precision and Recall and N was empirically fixed in a maximum of 250.

4. **Semantic based CE**

   As described in Section 3.7.4, the semantic representation of each sentence detected as relevant by the previous phase is used in a graph-based algorithm to create a candidate similarity matrix. Then, candidate summary sentences are selected taking into account three criteria: Relevance (with respect to the query or any other criteria), Density and cohesion, and Anti-redundancy. The same S and R factor 15% were applied to prevent overlapping and to remove not relevant candidates for both approaches (see Section 3.7.4 for details about the graph-based algorithm).

   In both *SEMsum* 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 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.

5. **Basis SC**

   While in *SEMsum* the SC input are those sentences selected by the CE component. In the *LEXsum* approach, the input are those sentences detected as relevant by RID. Then, SC is applied to obtain the summaries. Summary TUs are selected by relevance until the desired summary size is achieved. For each selected TU it is checked if the previous sentence in the original document also appears as a candidate, in that case both are added to the Summary in the order they appear in the original document.

Analysis of the results

The aspects to be analyzed in our participation in DUC 2006 were how well our different approaches performed in both detecting the most relevant sentences answering a specific user need and producing a non-redundant, cohesioned text. For that reason, analysis of the results focusses on the scores assigned by the NIST assessors to content responsiveness, and non-redundancy linguistic quality aspects. As said at the beginning of this chapter, each summary aspect was manually evaluated as 1: 'Very Poor', 2: 'Poor', 3: 'Acceptable', 4: 'Good', and 5: 'Very Good'.

Table 6.4 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 are shown 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 *SEMsum* approaches have a similar behavior, both obtaining an acceptable performance (around 3) in all aspects. Moreover, both of them perform especially well in non-redundancy (around 4). In contrast, *LEXsum* obtains only 2,43 in non-redundancy and referential clarity.

	Q1 Grammaticality		Q2 Non-redundancy		Q3 Referential clarity		Q4 Focus		Q5 Structure & coherence	
		Mean		Mean		Mean		Mean		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 6.4: FEMsum linguistic quality scores by approach, as well as the mean of the 34 participant systems obtained in the associated subset of summaries.

Content based responsiveness evaluates the amount of summary information that helps to satisfy the information need. The first column in Table 6.5 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); and the baseline (2,04). It can be observed that FEMsum is somewhat (0,04) above the participant mean (2,56).

To analyze the performance of each approach Table 6.6 shows in the first column the DUC participant score mean in summarizing the set of document clusters we assigned to each of our approaches. The second column presents the standard deviation. The third one gives the score obtained by our approaches. The fourth column shows the distance to the mean and the last column the ranking position.

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 6.5: DUC FEMsum manual content responsiveness score.

Being among the best participants, SEM150 is above the mean in 0.37, obtains an acceptable performance in content responsiveness (2.92) and ranks in a fourth position.

	Mean(1-35)	StDev	FEMsum	Mean Distance	Ranking
LEX	2.36	0.36	2.29	-0,07	14/35
SEM150	**2.55**	0.33	**2.92**	**0.37**	**4/35**
SEM250	2.58	0.33	2.53	-0.05	22/35

Table 6.6: DUC content responsiveness scores by approach.

Table 6.7 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 SEM250 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. In addition, SEM250 produces less 'Very Poor' summaries (6.7%) than LEX (14%).

	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 6.7: DUC content responsiveness scores distribution by approach

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).

The second manual content summary evaluation method 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 examples evaluated for each approach is not enough to analyze their performance according to this second perspective.

Finally, Table 6.8 presents the ROUGE-2 (R2) and ROUGE-SU4 (RSU4) scores obtained by the best ROUGE system (24), FEMsum (19), and the baseline (1). Differences between the ROUGE scores obtained 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.

	Best(24)	FEMsum(19)	Baseline(1)
R2	0.095	0.076	0.050
RSU4	0.155	0.131	0.098

Table 6.8: DUC ROUGE measures when considering 4 manual summaries as references.

6.1.4 SEMsum and SVM based approaches evaluated with autoPan

This section presents a first experiment in using SVM to detect relevant information to be included in a query-focused summary. Several classifiers are trained using pyramids of Summary Content Unit (SCU)s information.

In (Fuentes et al. 006b) SVMs are tested on two systems that participated in DUC 2006 contest. The performance of the new approaches is compared with the original systems using the DUC 2005 corpus as test data. For evaluation purposes, the automatic method based on pyramid data, AutoPan, presented in Section 6.1.1 is used.

Following the work of (Hirao et al. 2003), we envision the extraction of important sentences as a binary classification problem, where a sentence is either apt or not suitable for inclusion in a summary. Not only SVMs can be trained for this classification problem, but also they can rank the candidate sentences in order of relevance (Kazawa et al. 2002).

The analyzed approach consists in processing the set of documents, extract a set of features, and using the trained SVMs to rank the sentences and check them for redundancy in order to generate the final summary.

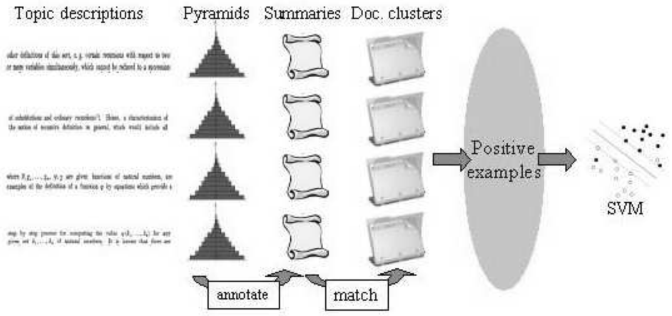


Figure 6.7: Procedure to train a Support Vector Machine to rank sentences.

Figure 6.7 illustrates the procedure to train a SVM. To do this, it is necessary to first build a training corpus. We used the DUC 2006 dataset, including topic descriptions, pyramid evaluations as annotated during the DUC 2006 manual evaluation, peer and manual summaries, and document clusters. From all these data, the training set is generated in the following way: first of all, the sentences in the original documents are matched with the sentences in the summaries as proposed by (Copeck and Szpakowicz 2005). Next, all document sentences that matched a summary sentence containing at least one SCU are extracted.

Every sentence from the original documents matching a summary sentence that contains at least one SCU is considered a positive example. To deal with the query-focused MDS the qMDS FEATURE EXTR component has been implemented with the SVM FEATURE EXTR.

As detailed in Section 3.6.2, the features extracted by the SVM FEATURE EXTR can be classified into three groups: those extracted from the sentences, those that capture a similarity metric between the sentence and the topic description, and those that try to relate the cohesion between a sentence and all the other sentences in the same document or collection.

Built in this way the training set contains only positive examples. The lack of negative examples is addressed as detailed in Section 3.7.2. The SVM CE uses the MC algorithm with positive, unlabeled data, and a small set of negative seeds.

The following settings can be varied during the experiments:

- Concerning positive examples, two different sets have been considered: those obtained by applying (Copeck and Szpakowicz 2005)'s proposal, and those obtained by matching document sentences to manual summaries.

- Concerning the kernel function of the SVM, we have experimented both with polynomial kernels (lineal, quadratic and cubic) and with Radial Basis Function (RBF) kernels.

- Concerning the MC algorithm, two procedures to choose the initial set of "strong negative" examples have been considered: a modification of the algorithm by (Yu, Han, and Chang 2002), to make it possible to handle non-binary features, and manually choosing a small set of negative examples from the unlabeled instances. In this last case, for each cluster, we took as negative sentences those found in the two or three automatic summaries with lowest manual Pyramid scores. In average, 11.9 sentences were selected for each document cluster as negative examples. In a similar way as with the positive sample, the selection of the negative sample could also be done automatically. However, it is not analyzed in this section.

For training, there were 6,601 sentences from the original documents, out of which around 120 were negative examples and either around 100 or 500 positive examples, depending on whether the document sentences had been matched to the manual or the automatic peer summaries. The rest were initially unlabeled.

The SVM CE has been instantiated by the *MLsum(MDS,Q)*, see the experiments reported in Section 6.1.4.

Evaluation Framework

To evaluate the performance of the SVM in detecting relevant information in query-focused MDS task, the DUC 2005 data has been used. The performance of each system is evaluated automatically using the ROUGE and AutoPan metrics (described in Section 6.1.1).

With respect to ROUGE measures, as in the last DUC competitions, both ROUGE-2 (R2) and ROUGE-SU4 (RSU4) were used to rank automatic summaries.

In this experiments, to compute the AutoPan score, we used the modified pyramid score computed from the peer annotation in DUC 2005 (AutoPan2 in Section 6.1.1). This score is a ratio of the sum of weights of the SCUs found in the peer (OBServed) to the sum for ideal summary (MAXimum). In the score used, MAX is computed using the average number of SCUs that were found in the seven human model summaries in the corresponding pyramid. Like recall, it indicates the proportion of the target highly weighted SCUs that were found in the peer.

The test corpus consists of 20 clusters manually evaluated in DUC 2005 with the pyramid method. Features of each sentence from each cluster were computed as described in Section 3.6.2. For complex questions the sentence splitting process presented in Section 3.5.1 was applied. This process creates new sentences when a conjunction that joins two words with the same POS is found.

The DUC 2006 systems evaluated are (Alfonseca et al. 2006), UAM-Titech06, and *SEMsum*.

Both evaluated systems are described bellow:

- The UAM-Titech06 system is a summarization system focused on producing coherent summaries. To that purpose, questions are divided into subquestions from which aims are identified, and separate mini-summaries are produced for each of the aims. These are later merged together in a final summary.

  The main steps in processing a document collection are the following:

  - Linguistic processing of the query and the original documents.
  - Identification of the aims of the questions, and background knowledge. For example, possible aims are *advantages, disadvantages, problems, legal privileges*, etc. of a certain status, specified by a set of background terms extracted from the topic description.
  - Automatic collection of separate document sets for each of the aims, and sentence ranking according to similarity to each set of documents.
  - Multi-summary generation, by choosing the top-ranked sentences from each of the ranks (with no repetitions).
  - Multi-summary merging in a single summary, generating a small introduction to each to indicate its focus.

- In contrast to the *SEMsum* approach evaluated in Section 6.1.3, in the evaluated *SEMsum* variant, NERC are computed, N is not fixed in the RID component, and in order to select the candidate sentences in the CE component the Relevance criteria is with respect to the query, the score given by the RID component is taken into account, and in the Candidates Selector procedure S is set to 15% of and R to 10%.

Analysis of the results

Table 6.9 is divided in two main parts: The first line shows the results obtained by the two original systems, UAM-Titech06 and *SEMsum.* In the second part of the table presents the results obtained when substituting the initial ranking step of the original approaches with different SVMs. Three different aspects has been studied:

- The use of positive examples: extracted from automatic summaries or manual summaries.
- The use of seed negative examples: manually annotated or automatically annotated.
- The use of different Kernels: RBF or Polynomial.

Some observations on the results are the following: firstly, concerning the different configurations of the SVM, some trends can be found:

Seed		UAM-Titech06			SEMsum		
negatives	**Kernel**	**R2**	**RSU4**	**autoPan**	**R2**	**RSU4**	**autoPan**
			Original Systems				
		0.048	0.105	0.052	**0.077**	**0.136**	0.066
			SVM (positive examples from peer summaries)				
Annotated	RBF	**0.071**	**0.131**	**0.072**	0.066	0.126	**0.069**
	Poly	0.062	0.119	0.064	0.061	0.118	0.052
Automatic	RBF	0.036	0.089	0.024	0.052	0.106	0.035
	Poly	0.055	0.113	0.058	0.058	0.117	0.056
			SVM (positive examples from manual summaries)				
Annotated	RBF	0.025	0.075	0.024	0.046	0.101	0.020
	Poly	0.046	0.102	0.053	0.043	0.098	0.024
Automatic	RBF	0.018	0.063	0.009	0.045	0.106	0,018
	Poly	0.038	0.087	0.021	0.044	0.099	0.028

Table 6.9: ROUGE and AutoPan results on the original UAM-Titech06 and SEMsum or when using several SVMs.

- SVMs trained using the set of positive examples obtained from the pyramid data consistently outperform SVMs trained using the examples obtained from the manual summaries. This may be due to the fact that the number of positive examples obtained from manual summaries (on average 12.75 per cluster) is smaller than from SCUs (on average 48.9).

- Generating automatically a set with seed negative examples for the MC algorithm, as indicated by (Yu et al. 2002), usually performs worse than using a set of seed negative examples selected manually from the SCU annotation. This may be due to the fact that its quality is better, even though the amount of seed negative examples is one order of magnitude smaller in this case (11.9 examples in average).

- The best results are obtained when using a RBF kernel, while previous summarization work (Hirao et al. 2003) uses polynomial kernels.

Concerning the two tested systems, UAM-Titech06 would be ranked in the middle zone among the participants in DUC 2005 in terms both of AutoPan and ROUGE. As can be seen in Table 6.9, a large gain can be obtained when combined with SVM, reaching very good ROUGE results, and attaining the best AutoPan result of all the systems evaluated in this Section (0.072). 0,081 is the score of the top AutoPan system (Daumé III and Marcu 2005), it also scored highest among DUC 2005 participant systems for responsiveness. UAM-Titech06's performance varies largely depending on the particular SVM used, probably due to the fact that the system just chooses the top-ranked sentences from the SVM output, so its output completely depends on the sentence rank received.

While UAM-Titech06's performance can be improved by using SVM, it can not be said that the substitution of *SEMsum* PR RID component for the SVM ranks constitute an increase in performance. On the one hand, *SEMsum* AutoPan score is better when using SVM (0.069) than in the original (0.066). However, this difference is not significant. On the other hand, in terms of ROUGE, the original *SEMsum* (0.077, 0.136) obtains better scores. It is ranked among the best DUC 2005 participant systems. The best participant (Ye et al. 2005) has a R2 score of 0.078 (confidence interval [0.07388 – 0.08075]) and an RSU4 score of 0.139 (confidence interval [0.13534 – 0.14264]), when evaluated on the 20 clusters used here. The difference between the scores obtained by the best system and those obtained by *SEMsum* are not statistically significant.

6.1.5 SEMsum at DUC 2007

The FEMsum approach for DUC 2007 main task (Fuentes et al. 007a) was based in the DUC 2006 best participation approach described in Section 6.1.3. The most important modification has been carried out in the RID component. For the DUC 2007 participation we implemented a TU re-ranking algorithm. As said in Section 3.6.3, TU Ranking module can be used in addition to the PR Software, JIRS in this experiments.

Analysis of the results

Ten NIST assessors wrote summaries for the 45 topics in the DUC 2007 main task. Two baseline summarizers were included in the evaluation: The first baseline (1) returns all the leading sentences (up to 250 words). The second baseline (2), CLASSY04 (Conroy et al. 2004), is an automatic summarizer that ignores the topic narrative but that had the highest mean Summary Evaluation Environment (SEE) coverage score in the MDS DUC 2004 Task 2.

NIST received submissions from 30 different participants for the main task. All summaries were truncated to 250 words before being evaluated.

Table 6.10 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). The last row presents the mean of FEMsum linguistic quality aspects and the system participant mean.

As in DUC 2006, acceptable quality summaries are obtained. The best score is achieved in non-redundancy aspect (3.71).

Content based responsiveness scores the amount of summary information that helps satisfy the information need. First column in Table 6.11 shows the responsiveness mean score obtained by: Humans (4.71), the best system (3.40), FEMsum (2.93), the baseline (1.87 and 2.71), and the participant mean score (2.64) and the last column shows the ranking position. It can be

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

observed that FEMsum scores above the mean and ranks in the eight position.

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	

Table 6.11: Content responsiveness score and mean distance for human, the best system, our submission and the baseline.

Finally, the *SEMsum* evaluated approach ranks in a 9th position when evaluated with BEs (score of 0.058, 0.008 above the mean). The baseline 1 ranks in the 29th position and the baseline 2 in the 19th.

6.2 Query focused Multi-Document Summarization update task

The DUC 2007 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 2007 task. For each topic, the documents are ordered chronologically and then partitioned into 3 sets, A, B, 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.

Most of the components used in FEMsum instantiation studied in this section, *SEMsum (Update)*, are the same as those in the instance evaluated in Section 6.1.5. Section 6.2.1 presents the modifications made to some of those components. And Section 6.2.2 analyzes the results obtained in the DUC 2007 manual evaluation.

6.2.1 Modifications with respect to DUC 2007 SEMsum

Figure 6.8 shows the FEMsum components used to participate in the DUC 2007 update task. The instantiated LP, QP and RID components are those used by *SEMsum* in the DUC 2007 main task, experiments reported in Section 6.1.5. The modifications carried out to some components to deal with this new task are presented in the following sections.

Semantic based CE: Candidate Selector

As described in Section 3.7.4, the CE requires two modules, the Similarity Matrix Generator (SMG) and the Candidates Selector (CS). This component requires document TUs (sentences in this instantiation) enriched with semantic information. Similarities among sentences are computed by SMG and used by the CS component.

The first change made in this component regarding the one used in DUC 2006 and the main DUC 2007 task is that in the previous approach the score assigned to each candidate sentence was computed using only the semantic similarities between the sentences coming from the RID component without taking into account the relevance assigned by JIRS. For the update task we have linearly combined this semantic similarity score with another one coming from the JIRS score. The way we have computed this last 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 anti-redundancy process. This new 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. The new approach is applied to iterations 2 and 3. In both cases a previous set of sentences are assumed to be known by the user (sentences from the A set of documents in the second iteration and A+B in the third). After the first and second iterations an additional antiredundancy step is performed for

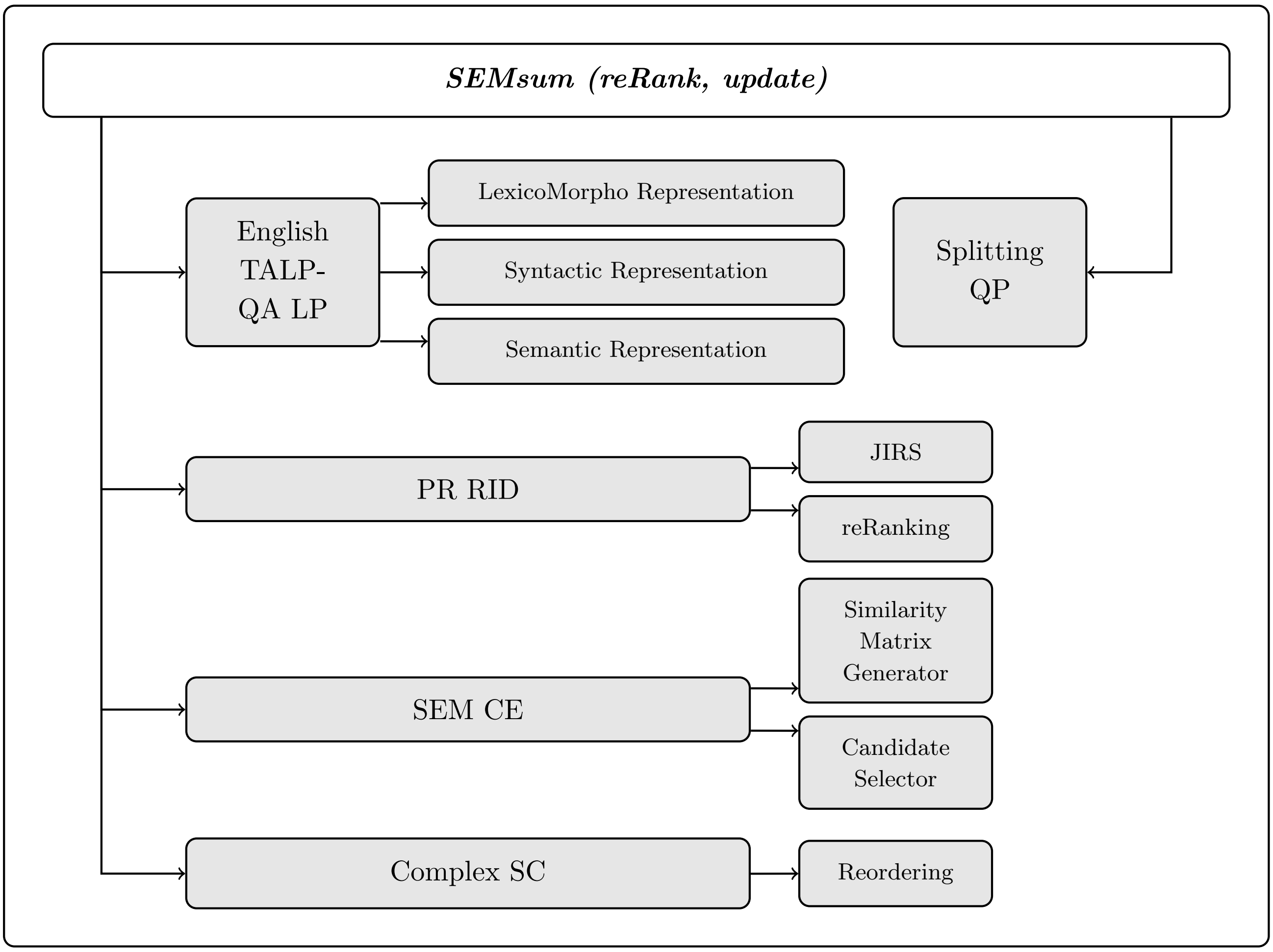


Figure 6.8: DUC 2007 FEMsum architecture for the update task.

preventing the duplication of information (see more details in Section 3.7.4).

Complex SC (Reordering)

The instantiated Reordering SC is detailed in Section 3.8.1. For the update task, instead of selecting the input TUs by relevance until reaching the summary size, we have introduced a new component for reordering the already selected sentences in a way of increasing cohesion, (Althaus et al. 2004).

6.2.2 Analysis of the results

Ten NIST assessors wrote summaries for the 10 topics in the DUC 2007 update task. 4 human summaries were written for each subset (A, B, C).

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.

Two baseline summarizers were included in the evaluation: The first one, identified with 35, returns all the leading sentences (up to 250 words), and the second one, identified with 58, CLASSY04 (Conroy et al. 2004) that uses HMM with signature terms. CLASSY04 choose sentences 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.

Table 6.12 shows the results obtained in the update task. The first column presents the responsiveness score obtained by FEMsum, the second column presents the ranking, the third one the participation mean. The three last columns show the scores obtained by the best system, identified as 40, and the results obtained by the two baselines.

	SEMsum (Update)	Rank	Mean	Best (40)	Baseline1 (35)	Baseline2 (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

Table 6.12: *SEMsum (Update)* responsiveness evaluation, 22 participants.

Each row in Table 6.12 presents the performance obtained summarizing the corresponding set of (A, B and C). The last row gives the mean performance. Our system performs somewhat under the mean, but obtaining always better results than the Baseline1.

6.3 Conclusions

We have presented a method to automatically evaluate different FEMsum instantiations to deal with the real-world complex question answering task proposed at DUC. This method uses the pyramid information manually created.

We believe that AutoPan a part to be useful to evaluate automatic systems in a faster and less costly way, it can be also interesting as a tool for human annotators. When evaluating sets of summaries, it is difficult for humans to keep constant criteria along the complete set. AutoPan can give a homogenous starting point that only needs to be corrected by the annotator. There is still work to be done on our tool, but being AutoPan a method which uses such shallow linguistic information (only stemming), the fact that correlation with human judgment is significant is encouraging.

The AutoPan performance highly depends on the number of summaries used in the pyramid construction process. We have observed a decrease in the method utility when using less summary references. For that reason, the method has been applied only to DUC 2005 data where pyramids were created from seven reference summaries. In contrast, in DUC 2006 and 2007 only 4 references were used in the pyramid creation process.

A first analyzed FEMsum MDS approach, *QAsum*, generates a set of simpler queries from the complex one. Then, part of a QA system, TALP-QA, is used to detect relevant sentences. The semantic representations of all the relevant sentences are contrasted in order to avoid redundancy. *QAsum* prototype was evaluated at DUC 2005 and some malfunctions were detected. A new prototype was evaluated using AutoPan. It was observed that the performance of TALP-QA PR decreases when working in the DUC 2005 task.

With the aim of evaluating several aspects of new FEMsum approaches we submitted different kinds of automatic summaries in the same run to be evaluated at DUC 2006. Two different types of summarization approaches have been examined. A first one, based on lexical information, *LEXsum*, and a second one, that uses semantic information, *SEMsum* in order to avoid redundancy and to improve the cohesion of the resulting summary. One of the changes with respect to *QAsum* is that JIRS, a PR software, is used instead of the TALP-QA PR to instantiate the RID component. And *SEMsum* deals with redundancy in CE instead of in the SC component.

The aspects to be analyzed in our participation in DUC 2006 were how well new approaches performed in both detecting the most relevant sentences answering a specific user need and producing a non-redundant, cohesioned text. The experiments show that using semantic information significantly increases the performance when dealing with written news articles. The best *SEMsum* configuration was evaluated in DUC 2007 ranking in the top ten for responsiveness and producing acceptable non-redundant summaries.

Some experiments are presented to take advantage of the annotations created during DUC conference for training automatically text summarization approaches using ML techniques. Different possibilities for applying DUC information in training SVMs to be used as RID component have been studied. The experiments have provided some insights on which can be the best way to exploit the annotations. On the one hand, the positive examples obtained from the annotations of the peer summaries are more useful than those obtained from the annotations of model summaries. That is probably due to the fact that most of the peer systems are extract-based, while the manual ones are abstract-based. On the other hand, using a very small set of negative example seeds seems to perform better than choosing automatically the negative examples as the ones that are more different to the positive instances.

The best SVM obtained has been able to produce a very large improvement on the UAM-Titech06 system, but not significative improvement, in terms of the AutoPan evaluation metric, when applied to the *SEMsum* summarization system.

The best *SEMsum* DUC main task approach 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 specific reordering algorithm was used. Our system performed somewhat under the mean, but obtaining always better results than the baseline that consists in returning all the leading sentences (up to 250 words). For the update task linguistic quality aspects were not evaluated, for that reason, we are not able to evaluate if applying the specific reordering affects in the performance of the system.