Chapter 7

Multi-source FEMsum

This chapter studies the FEMsum performance when dealing with a new query-focused MDS task within CHIL project framework. The task consists in summarizing an oral presentation in order to give answer to a user need expressed as a list of keywords. Given a query or list of relevant terms (e.g. "automatic speech recognition", "perplexity measures", "N-gram language model") and a set of documents, the summarizer is required to return a fixed-length extract of relevant segments (100 words) from multiple documents to answer a set of queries (relevant terms).

The goal of the MDS focused by queries task studied in this chapter is to answer a query by summarizing documents of different natures. In this task, as in the SDS task studied in Chapter 5, we focused on the CHIL lecture scenario. The lecture takes place in a Smart Room that has access to different types of documents related to the oral presentation to be summarized. Concretely, a multi-document set may consist of documents produced from different media regarding a specific lecture, such as:

- The scientific paper(s) to which the lecture refers.
- The manual transcript of the audio recording.
- The text of the corresponding presentation slides.
- The author notes, if available.

Studies such the one presented in (Shriberg 2005) show that oral communication is harder to process than written text. For that reason, we propose using a MDS approach capable of handling documents from different media types to summarize the content of scientific oral presentations. Combining documents from different media can help counteract not only the difficulties in processing oral communication, but also those errors introduced by ASRs.

Next section presents the CHIL evaluation framework. Section 7.2 briefly describes five different approaches that have been manually and automatically evaluated. Section 7.3 analyzes their performance in the task proposed in this chapter. Then, in Section 7.4 the results obtained by *SEMsum* and *LEXsum* approaches in this new query-focused MDS task are contrasted with the results obtained in the DUC task previously studied in Chapter 6.

7.1 Evaluation framework

The final objective of the reported research is to integrate a summarizer in a Smart Room. As said in Chapter 5, the Smart Room is an intelligent space equipped with several distributed cameras and microphones. Moreover, the room has access to the digital material used by the speaker. In this scenario, the summaries to be presented to the user would be of different types: fragments of the image/audio file containing the most relevant information, pieces of the digital material used or cited by the speaker, or voice synthesized from the textual summary. Due to the fact that it is possible to produce non-textual summaries it was decided to start by evaluating only textual summary content aspects.

Evaluation and Language resources Distribution Agency (ELDA)[1], in charge of the creation of the CHIL test corpora, selected 10 seminars from the ones recorded at Karlsruhe University, at the Interactive Systems Laboratories (ISL)[2]. Table 7.1 presents an example of the ISL topics.

Seminar ID	**Topic**
ISL_20031111	Robustness through articulatory features
ISL_20041112_A	Speech translation
ISL_20041123_A	Grapheme based speech recognition
ISL_20041123_E	ISL meeting transcription system
ISL_20050112	Blind segment of acoustic signal

Table 7.1: Example of ISL seminar topics.

Different sorts of documents (4 on average) were collected for each of these speech events (technical seminars). For instance, besides the manual transcription of the seminar, additional documents from scientific publications, conference papers, and presentation slides related to the seminar, if available. All documents were converted into plain text. For each set of documents, two queries were generated according to the seminar topic and the documents content. An example of queries is given in Table 7.2.

[1] http://www.elda.org/
[2] http://isl.ira.uka.de/

Seminar ID	**Queries**
ISL_20031111	*Q1*: Articulatory features + automatic speech recognition + HMM *Q2*: Large vocabulary continuous speech recognizer + LVCSR
ISL_20041112_A	*Q1*: Statistical Machine Translation + Noisy Channel Paradigm *Q2*: Phrase-based spoken language translation + automatic speech recognition + word lattice
ISL_20041123_A	*Q1*: Multilingual grapheme based speech recognition + poly-grapheme clustering *Q2*: Pronunciation dictionary + classification and regression trees + CART

Table 7.2: Example of queries for three of the ISL seminar topics.

For each of two generated queries of each subset of documents, three human annotators were asked to create extracts with a length of approximately 100 words by concatenating relevant segments from multiple documents to answer the given query. The generated extracts were used as reference summary models for the automatic evaluation. Table 7.3 presents three assessors models produced for the *Speech Translation* topic, with the query: *Statistical Machine Translation + Noisy Channel Paradigm.*

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

In addition to the automatic evaluation, a manual content responsiveness evaluation was carried out at UPC, along the same lines as in DUC. Concretely, 20 assessors were asked to score each automatic and manual summary in terms of summary responsiveness content. Figure 7.1 shows the interface used by the UPC assessors in the evaluation process. The top part of the screen presents the list of relevant terms to be taken into account in the summary production *Statistical Machine Translation + Noisy Channel Paradigm.* As in DUC, UPC assessors were asked to assign an integer between 1 (least responsive) and 5 (most responsive) with respect the three reference summaries created at ELDA (at left site in Figure 7.1). Human evaluators were also asked to score the quality of the human summaries taking into account the other two human models as reference.

The translation model used is the CMU Statistical Machine Translation toolkit. The one best sentence inputted is converted into a single path lattice first. Statistical machine translation work is based on the noisy channel paradigm. The noisy-channel approach is a basic concept of SMT that was adapted to the problem of disfluency cleaning, by assuming that "clean" i.e. fluent speech gets passed through a noisy channel. The modelling task is to build statistical models which capture the characteristics of the translation and the target language.
In conventional speech translation systems, the recognizer outputs a single hypothesis which is then translated by the Statistical Machine translation (SMT) system. This approach has the limitation of being largely dependent on the word error rate. A translated sentence from English to French is garbled or is encoded by some noisy channel and we need to model this noisy channel in some statistically meaningful way . The cleaning component is based on a noisy channel approach, a basic concept of SMT that is adapted to the problem of disfluency cleaning, by assuming that fluent speech gets passed through a noisy channel.
While speech recognition emerged to be rapidly adaptable to new languages in large domains, translation still suffer from lack of large parallel corpora for statistical machine translation. Statistical machine translation work is based on the noisy channel paradigm. The model to describe this noisy channel needs to provide us with a probability of an English target sentence given a French source sentence. In SMT we use alignments to establish correspondences between the positions of the source and the target sentences. In conventional speech translation systems, the recognizer outputs a single hypothesis which is then translated by the SMT system.

Table 7.3: Example of the three reference summary models of one of the 20 summaries to be evaluated.

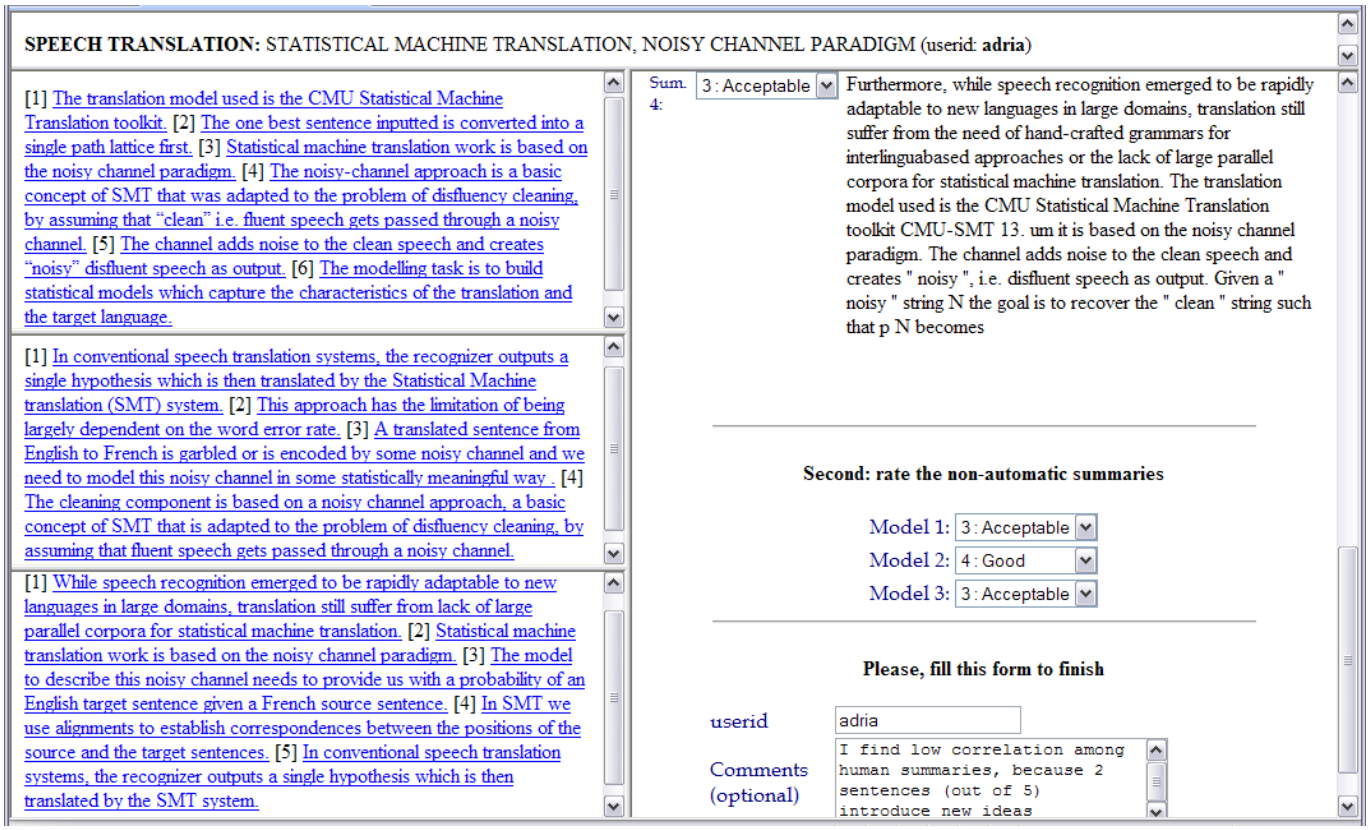

Figure 7.1: CHIL manual evaluation web interface.

7.2 Configurations of the evaluated approaches

The approaches that participated in the CHIL evaluation were:

- SDS. We assume that adding textual information helps when summarizing spontaneous speech, for that reason the PostSeg lexical chains based SDS, *LCsum (PostSeg)* (described in Section 5.3), has been used as a baseline. This approach extracts segments of about 30 words only from the transcription and has been adapted to be query-driven by increasing the weight of the lexical chain members that appear in the query.

- LEX. As described in Section 6.1.3, *LEXsum* only uses lexical information. Segments from all documents are candidates to appear in the summary.

- SEM. Due to the fact that textual transcriptions from spontaneous speech are often ill-formed and they not always follow the written syntactic rules, transcription segments have not been considered as summary candidates in *SEMsum*, described in Section 6.1.3. Having

a small number of documents to be summarized, we decided to use the not fixed default value of N. All the TUs selected as relevant by some JIRS execution were considered as RID output. The number of TUs detected as relevant ranges from 67 to 257 TUs (186,75 in average). The CE module settings are: 15% of S and 10% of R. Finally the Basic SC component was instantiated to produce the final summary.

- LEXnoT. It is as *LEXsum* approach, but without considering the transcription segments as summary candidates.

- +www. In the CHIL corpus, the number of documents to be summarized is smaller than in the DUC one. However, since a lot of scientific information is available online, it would be interesting to evaluate the UAM-Titech06 system briefly presented and evaluated in Section 6.1.4, which also participates in the DUC 2006 evaluation. This system takes into account background information related to the query from the World Wide Web in order to produce the summaries. This system obtains similar results to general FEMsum participation in DUC contest.

7.3 Analysis of the results

Table 7.4 shows the ROUGE metric results when comparing 20 extract-based summaries of a set of documents related to scientific presentations, against three human-created summaries. Best ROUGE values are shown in bold. All the evaluated approaches perform better than the baseline, SDS. Looking at the ROUGE measures, it is difficult to determine whether LEXnoT is better or not than +www.

Until now, for the new summarization task proposed in CHIL, it is not possible to show good and consistent correlation between ROUGE measures and human evaluations. To have manual evaluations is the first step to try to solve that. However, the amount of human evaluations is not enough to conclude what is the ROUGE measure that better correlates with human judgments to be used to score system performance.

Looking at the content responsiveness results, in Table 7.5, we see that LEXnoT obtains the best mean score (2.025), while +www and SEM obtain the same score (1.800). The lower score obtained by a MDS approach is the one obtained by LEX (1.775). That means that better mean performance is obtained when not using the transcription as part of the output.

To find out how effective is each approach Table 7.6 shows the percentage of summaries classified by score. On the one hand, although SEM mean is the same as +www (1.8 in Table 7.5), Table 7.6 shows that the percentage of summaries considered as 'Acceptable' or 'Good' is higher in SEM (20% + 5%) than in +www (15% + 5%). On the other hand, LEX with a lower mean

	SDS	LEX	LEXnoT	+www	SEM
ROUGE-1	0.293	0.309	0.312	**0.333**	0.323
ROUGE-2	0.060	0.092	**0.102**	0.089	0.073
ROUGE-3	0.029	0.056	**0.064**	0.052	0.032
ROUGE-4	0.019	0.043	**0.050**	0.043	0.021
ROUGE-L	0.256	0.272	0.279	**0.289**	0.280
ROUGE-W1.2	0.089	0.098	0.100	**0.104**	0.098
ROUGE-S1	0.057	0.088	**0.097**	0.087	0.067
ROUGE-S4	0.064	0.089	**0.095**	0.094	0.073
ROUGE-S9	0.069	0.095	0.102	**0.103**	0.083
ROUGE-SU1	0.136	0.162	**0.169**	0.168	0.152
ROUGE-SU4	0.102	0.126	0.132	**0.134**	0.115
ROUGE-SU9	0.090	0.116	0.122	**0.124**	0.105

Table 7.4: ROUGE measures when considered 3 manual summaries as references.

M1	M2	M3	SDS	LEX	LEXnoT	+www	SEM
3.625	3.400	3.375	1.250	1.775	**2.025**	1.800	1.800

Table 7.5: Responsiveness considering 3 human models when evaluating automatic summaries and 2 when evaluating human summaries.

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

Table 7.6: Responsiveness scores distribution by automatic system.

score (1.775) obtains better results than SEM or +www in terms of percentage score with a 35% of acceptable summaries.

7.4 Contrasting the results obtained in different tasks

The aim of this section is to contrast the performance of *LEXsum* and *SEMsum* when dealing with two different query-focused MDS tasks. Although both DUC and CHIL tasks are query-focused MDS, they differ in the following aspects:

- the type of query (complex vs. list of relevant terms)
- the number of documents to be summarized (25 vs. 4)
- the input media (well written text vs. raw text coming from: spontaneous speech transcripts and documents related to an oral presentation)
- the input domain (news vs. scientific)
- the input genre (written journalism vs. scientific spontaneous speech, research papers, and author notes or slides)
- the output summary length (250 vs. 100)
- the size of the evaluation test corpus (50 vs. 20)
- the number of manual summary references (4 abstract-based vs. 3 extract-based)

We have contrasted the responsiveness score obtained in DUC 2006 evaluation (see Table 6.5) and the one obtained in the CHIL evaluation (see Table 7.5). The performance obtained from human assessors in DUC (4.75) compared with the one obtained from CHIL assessors (3.48), it seems that the task proposed in the CHIL framework is harder than the one in DUC. Nevertheless, although the score obtained by the best CHIL system (2.025), is lower than the best DUC system (3.08), the distance from the upper bound (human assessors) is smaller in the CHIL task (3.48, 2.035) than in the DUC challenge (4.75, 3.08). In fact, the score gap between human and system performance is smaller when summarizing scientific presentations than when considering a set of written news as input. This difference can be partly explained by the fact that less manual summaries were used as reference in the CHIL evaluation. While in DUC evaluation 4 manual summaries were always used as reference, in CHIL 3 references were used to evaluate system summaries and only 2 to evaluate human summaries.

Looking at the performance of the proposed approaches in the two different scenarios, we observe that LEX seems to be robust in terms of responsiveness. LEXnoT scores are 2.025 in CHIL (see Table 7.6) and 2.29 in DUC (Table 6.7). Given that LEX mainly uses the RID component, this means that the detection of relevant information for the summary is similar for both tasks: oral presentations and written articles. However, because of the genre of the documents to be summarized, the difficulty of the task, and that the linguistic processor tools

were trained for general domain, SEM versions obtain better results in DUC (2.53 and 2.92) than in CHIL (1.8). More efforts have to be done in order to set the appropriate parameters of the CE module used for the CHIL scenario.

7.5 Conclusions

This chapter shows that FEMsum is capable of dealing with different MDS summarization tasks by combining information from documents of different types, like manual transcriptions, slides, notes or well written text. FEMsum approaches also take into account the user needs as well as the particular features of the documents to be summarized. Within this framework, we have focused on summarizing written news articles and textual documents related to scientific oral presentations in response to a query.

Two different types of summarization approaches have been examined. *LEXsum*, based on lexical information, and *SEMsum*, that uses semantic information in order to avoid redundancy and to improve the cohesion of the resulting summary.

The experiments show two main facts. On the one hand, the approaches using only lexical information achieve similar performances in both scenarios: written news articles and scientific oral presentations. On the other hand, adding semantic information significantly increases the performance when dealing with written news articles. In contrast, there is room for improvement when adding semantic information when dealing with different sorts of documents from the scientific domain, most likely because the language processing tools were trained on a different domain. Moreover, oral presentation documents are less structured than formal text and less edited.