Chapter 5

Spontaneous Speech FEMsum

With the increasing importance of human-machine interaction, speech, as the most natural way of human communication, has become the core of many computer applications. Moreover, Automatic Summarization (AS) is of help in digesting the increasing amounts of information that reach us every day. For that reason, in the framework of the *Computers in the Human Interaction Loop (CHIL)*[1] project, we have investigated how the combination of speech and summarization could improve communication between humans and computers. In this project, human activities, intentions and interactions are studied to provide helpful services.

With the objective of allowing users to obtain information about previous events, in the CHIL project, FEMsum is integrated in a smart-room environment. The CHIL Smart Room is an intelligent space equipped with several distributed cameras and microphones enabling the possibility of multimodal processing. At the Universitat Politècnica de Catalunya (UPC), the Smart Room has been equipped with 85 microphones and 8 cameras. A CHIL scenario is a situation in which people interact face to face with other people, exchange information, collaborate to jointly solve problems, learn, or socialize, by using whatever means (speech/language, gestures, body posture, data in electronic format, slides, etc.) they choose. CHIL monitors the environment and provides useful services.

The Single Document Summarization (SDS) task studied in this chapter is focused in the scenario where an oral presentation has taken place, exploiting manual and automatic transcripts of the audio recording of the presentation. Given a transcript, three different summaries have to be produced: ultra-short (10-word), short (30-word) and long (70-word). The final output to be given to the user is the reproduction of the corresponding video and/or audio segments.

The framework presented in Section 5.1 was created to help in the evaluation of this new task. This evaluation framework has been used in two different experiments. First, in Section 5.2, the original textual news FEMsum instantiations presented in Chapter 4 are evaluated with

[1] chil.server.de/servlet/is/101/ (IST-2004506969)

automatic and manual transcripts as input. Second, in Section 5.3, adapted instantiations are evaluated to set up the best FEMsum configuration to summarize automatically transcribed oral presentations. The first experiment is part of the work reported in (Fuentes et al. 005a) and the second one was first published in (Fuentes et al. 005b).

5.1 Spontaneous speech evaluation framework

Evaluation of the FEMsum approaches has been carried out using the Translingual English Database (TED)[2] (Lamel et al. 1994). This corpus is composed by a set of audio recorded speeches from non-native English speakers presenting academic papers, in the EuroSpeech'93 conference, for approximately 15 minutes each. This can be classified as quasi-spontaneous speech. The presentations were performed live without reading from a script, but it is safe to assume that the speaker followed a self prepared high-level script. Furthermore, speakers were not interrupted during their presentation.

The corresponding paper was available for most of the presentations, so they were studied in order to try to use the valuable information they have. As detailed in Section 5.2.1 the title and keywords can be taken as good summary models. However, the abstract can not be considered as a good summary model for long summaries of oral presentations.

Additionally, the TED corpus contains manual transcripts for 39 of the presentations, 37 of which have their corresponding paper.

To evaluate the performance when summarizing scientific oral presentations, three different sets of documents from TED have been used to create three different gold standard corpora. Each one contains summary models of different lengths. Section 5.1.1 describes the manual production of summaries for the 39 manually transcribed presentations. Section 5.1.2 presents the process by which 29 of the 39 presentations have been automatically transcribed and manually summarized. To finish, Section 5.1.3 presents a corpus of 93 automatic transcripts to be used as a test corpus taking into account only paper summary models (title and list of keywords).

5.1.1 TED manual transcript gold corpus

Apart from the *abstract*-based paper summary models (title, list of keywords and abstract), *extract*-based gold standard summaries were manually created for each one of the 39 manually transcribed TED presentations. Summaries of about 10, 30 and 70 words were produced by manually extracting relevant chunks from each transcript.

Because of the lack of segmentation or punctuation when working with spontaneous speech documents, human assessors were asked to select and rank the most important chunks of text

[2] http://www.elda.org/catalogue/en/speech/S0120.html

from each document. Figure 5.1 shows the interface of the tool developed to make this annotation process easier. As summarized in the figure, the manual summary content selection process consists in four steps:

- First, chunks are selected and labeled with "1", "2" or "3", according to their relevance. An identifier is automatically assigned to each new chunk and *subst* is used if the annotator wants to indicate a set of equivalent chunks. That is useful for automatic summary selection. Furthermore, the tool allows establishing dependencies with other chunks when a chunk should appear only if another specific chunk appears. In Figure 5.1 the chunk *identified* by *4* *"it's been modified for this AUDETEL application"* has been assigned the lowest weight, *3* and this chunk can be *substituted* and depends on *3*, *"CELP"*.
- Second, the summary length is defined in words (right side of the figure) or percentage (left side).
- Third, the tool automatically proposes a summary chunk selection taking into account the *subst* and *weight* information.
- Fourth, the annotator supervises and changes the final chunk selection, if necessary.

Given the manual annotation in the first step, this process aims to be flexible enough to allow for the creation of gold standard summaries of different sizes automatically, if necessary. For the reported experiments, three summary test corpora were manually generated: 10-word, 30-word and 70-word summary length.

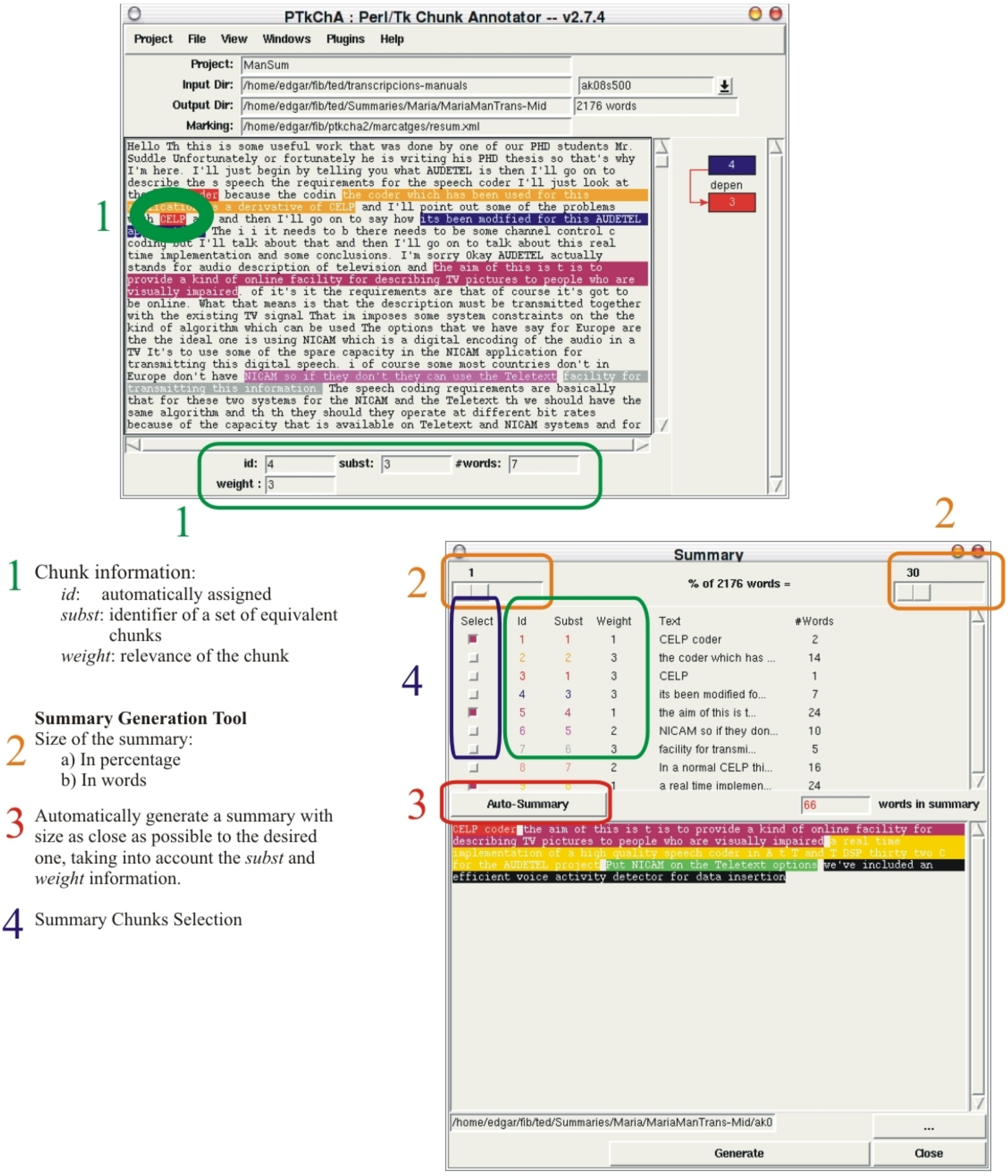
5.1.2 JANUS-ViaVoice automatic transcript gold corpus

To create an *extractive* gold standard, the voice files of 8 of the 39 presentations used in the previous section were automatically transcribed using the UKA JANUS (Lavie et al. 1994) system, one of the CHIL project speaker independent ASRs. JANUS has a Word Error Rate (WER) of 31%.

WER is the measure of the number of modifications required to transform the automatically recognized word sequence into the reference sequence, in terms of number of insertions (I), deletions (D) and substitutions (S). Being N the number of words in the reference, the formula is:

$$WER = (I + D + S)/N$$

Table 5.1 shows a small passage from one of the 8 automatic transcripts. As illustrated, the format of the transcripts follows several widely used conventions:



PTkChA : Perl/Tk Chunk Annotator -- v2.7.4
Project File View Windows Plugins Help
4
depen
3
1
1
2
2
2
3
4
Summary
Auto-Summary
...
Generate
Close

Project: ManSum

Input Dir: /home/edgar/fib/ted/transcripcions-manuals ak08s500

Output Dir: /home/edgar/fib/ted/Summaries/Maria/MariaManTrans-Mid 2176 words

Marking: /home/edgar/fib/ptkcha2/marcatges/resum.xml

```
Hello Th this is some useful work that was done by one of our PHD students Mr.
Suddle Unfortunately or fortunately he is writing his PHD thesis so that's why
I'm here. I'll just begin by telling you what AUDETEL is then I'll go on to
describe the s speech the requirements for the speech coder I'll just look at
the CELP coder because the codin the coder which has been used for this
application is a derivative of CELP and I'll point out some of the problems
with CELP and and then I'll go on to say how its been modified for this AUDETEL
application The i i it needs to b there needs to be some channel control c
coding but I'll talk about that and then I'll go on to talk about this real
time implementation and some conclusions. I'm sorry Okay AUDETEL actually
stands for audio description of television and the aim of this is t is to
provide a kind of online facility for describing TV pictures to people who are
visually impaired. of it's it the requirements are that of course it's got to
be online. What that means is that the description must be transmitted together
with the existing TV signal That im imposes some system constraints on the the
kind of algorithm which can be used The options that we have say for Europe are
the the ideal one is using NICAM which is a digital encoding of the audio in a
TV It's to use some of the spare capacity in the NICAM application for
transmitting this digital speech. i of course some most countries don't in
Europe don't have NICAM so if they don't they can use the Teletext facility for
transmitting this information. The speech coding requirements are basically
that for these two systems for the NICAM and the Teletext th we should have the
same algorithm and th th they should they operate at different bit rates
because of the capacity that is available on Teletext and NICAM systems and for
```

id: 4 subst: 3 #words: 7
weight : 3

1 Chunk information:
id: automatically assigned
subst: identifier of a set of equivalent chunks
weight: relevance of the chunk

Summary Generation Tool

2 Size of the summary:
a) In percentage
b) In words

3 Automatically generate a summary with size as close as possible to the desired one, taking into account the *subst* and *weight* information.

4 Summary Chunks Selection

1 % of 2176 words = 30

Select	Id	Subst	Weight	Text	#Words
☒	1	1	1	CELP coder	2
□	2	2	3	the coder which has ...	14
□	3	1	3	CELP	1
□	4	3	3	its been modified fo...	7
☒	5	4	1	the aim of this is t...	24
□	6	5	2	NICAM so if they don...	10
□	7	6	3	facility for transmi...	5
□	8	7	2	In a normal CELP thi...	16
☒	9	8	1	a real time implemen...	24

66 words in summary

```
CELP coder the aim of this is t is to provide a kind of online facility for
describing TV pictures to people who are visually impaired a real time
implementation of a high quality speech coder in A t T and T DSP thirty two C
for the AUDETEL project Put NICAM on the Teletext options we've included an
efficient voice activity detector for data insertion
```

/home/edgar/fib/ted/Summaries/Maria/MariaManTrans-Mid/ak0

Figure 5.1: Summary Generation Tool.

> my name is you can and i will present results from wizard of oz experiments
> performed at the centre for cognitive science university denmark to the
> experiments from a test part of the project for the dialogue project a spoken
> language dialogue systems which is carried out in collaboration two sentence
> the of the dialogue project is to develop two prototypes what he wanted to
> repeat to be somewhat constitution of pea one in both systems should be within
> the domain of danish domestic flight reservation and information tasks which
> it should be possible to observation of tickets changes of observations and
> information departure stressed conditions

Table 5.1: Example of content from an automatic transcript.

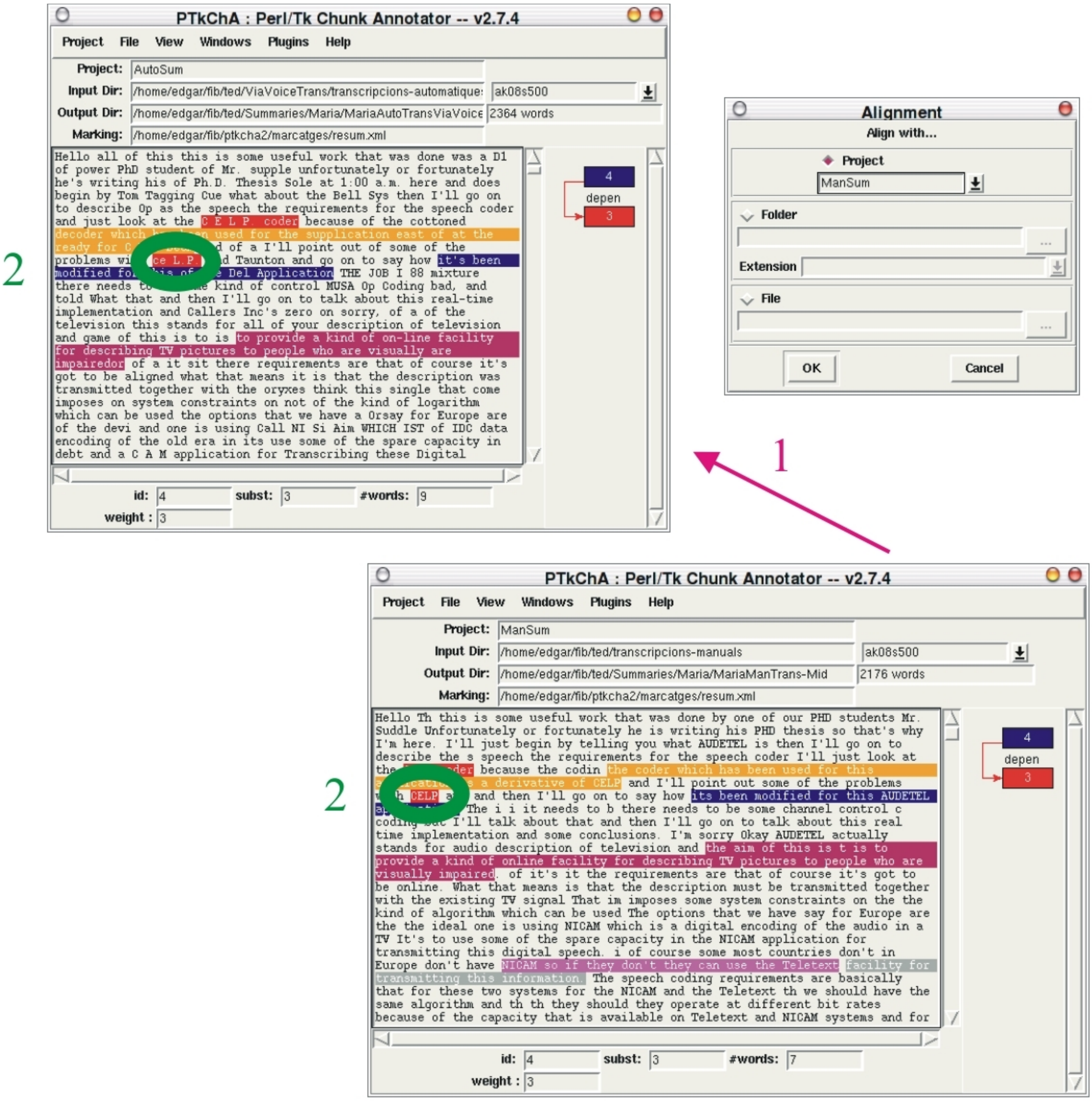
- All words are transcribed in lower case.
- No punctuation marks are used.
- Numbers are written alphabetically (e.g. “two”) not numerically.
- Speech noises and other outside noises are omitted.

As all CHIL ASRs were trained with 31 documents within the 39 in our evaluation framework, it was not possible to use JANUS or any other CHIL ASR to generate automatic transcripts for all 39 documents. 8 is a small number of examples to be statistically significant, for that reason, we decided to train the IBM software ViaVoice to generate 21 additional transcripts. In this case, the WER of the obtained automatic transcripts was about 38%. These automatic transcripts were generated by reading aloud the corresponding manual transcripts to the dictaphone, previously trained with the speaker’s voice.

The same assessors that created summary models for manual transcripts were asked to create equivalent models for automatic ones. However, since the low quality of automatic transcription makes the reading process difficult, we simplified the task of creating models, by previously aligning automatic transcripts with the chunks annotated in the corresponding manual transcripts. To produce model summaries for each document, assessors were asked to review, and, if necessary, modify, the aligned chunks. This implies that the assessor has some initial bias about the relevance of chunks. Depending on the quality of the automatic transcript, automatically aligned chunks were more or less modified. Figure 5.2 shows the process of generating new models by aligning and correcting chunks when necessary. Once chunks are corrected, extract-based summaries are procuced as with manual transcripts, detailed in Section 5.1.1.

5.1.3 JANUS automatic transcript gold corpus

To carry out a more extensive evaluation, 93 of the TED audio files, without manual transcription, were automatically transcribed using JANUS. As said before, JANUS was trained with



Project: AutoSum
Input Dir: /home/edgar/fib/ted/ViaVoiceTrans/transcripcions-automatique: ak08s500
Output Dir: /home/edgar/fib/ted/Summaries/Maria/MariaAutoTransViaVoice 2364 words
Marking: /home/edgar/fib/ptkcha2/marcatges/resum.xml

Hello all of this this is some useful work that was done was a D1 of power PhD student of Mr. supple unfortunately or fortunately he's writing his of Ph.D. Thesis Sole at 1:00 a.m. here and does begin by Tom Tagging Cue what about the Bell Sys then I'll go on to describe Op as the speech the requirements for the speech coder and just look at the C E L P. coder because of the cottoned decoder which h[illegible]en used for the supplication east of at the ready for C[illegible]d of a I'll point out of some of the problems wi[illegible] ce L.P. [illegible]d Taunton and go on to say how it's been modified fo[illegible]e Del Application THE JOB I 88 mixture there needs to[illegible]e kind of control MUSA Op Coding bad, and told What that and then I'll go on to talk about this real-time implementation and Callers Inc's zero on sorry, of a of the television this stands for all of your description of television and game of this is to is to provide a kind of on-line facility for describing TV pictures to people who are visually are impairedor of a it sit there requirements are that of course it's got to be aligned what that means it is that the description was transmitted together with the oryxes think this single that come imposes on system constraints on not of the kind of logarithm which can be used the options that we have a Orsay for Europe are of the devi and one is using Call NI Si Aim WHICH IST of IDC data encoding of the old era in its use some of the spare capacity in debt and a C A M application for Transcribing these Digital

id: 4 subst: 3 #words: 9
weight : 3

Alignment
Align with...
◆ Project ManSum
Folder
Extension
File

Project: ManSum
Input Dir: /home/edgar/fib/ted/transcripcions-manuals ak08s500
Output Dir: /home/edgar/fib/ted/Summaries/Maria/MariaManTrans-Mid 2176 words
Marking: /home/edgar/fib/ptkcha2/marcatges/resum.xml

Hello Th this is some useful work that was done by one of our PHD students Mr. Suddle Unfortunately or fortunately he is writing his PHD thesis so that's why I'm here. I'll just begin by telling you what AUDETEL is then I'll go on to describe the s speech the requirements for the speech coder I'll just look at the [illegible]der because the codin the coder which has been used for this a[illegible]s a derivative of CELP and I'll point out some of the problems w[illegible]h CELP a[illegible] and then I'll go on to say how its been modified for this AUDETEL ap[illegible] The i i it needs to b there needs to be some channel control c codin[illegible] I'll talk about that and then I'll go on to talk about this real time implementation and some conclusions. I'm sorry Okay AUDETEL actually stands for audio description of television and the aim of this is t is to provide a kind of online facility for describing TV pictures to people who are visually impaired. of it's it the requirements are that of course it's got to be online. What that means is that the description must be transmitted together with the existing TV signal That im imposes some system constraints on the the kind of algorithm which can be used The options that we have say for Europe are the the ideal one is using NICAM which is a digital encoding of the audio in a TV It's to use some of the spare capacity in the NICAM application for transmitting this digital speech. i of course some most countries don't in Europe don't have NICAM so if they don't they can use the Teletext facility for transmitting this information. The speech coding requirements are basically that for these two systems for the NICAM and the Teletext th we should have the same algorithm and th th they should they operate at different bit rates because of the capacity that is available on Teletext and NICAM systems and for

id: 4 subst: 3 #words: 7
weight : 3

1 Using the alignment plug-in, automatic transcripts are aligned with manual ones in order to automatically create chunk information for the former.

2 Matching chunks have the same information: span, weight, subst and dependencies.

Figure 5.2: Automatic transcript chunks aligned with chunks from the manual transcript.

31 of the 39 files with manual transcription, for that reason none of those 31 presentations was used for evaluation.

Due to the high cost of producing summary models manually, to obtain the summary models associated to each of these 93 presentations, we decided to use the abstract-based summaries found in the scientific paper associated to each oral presentation. The title, keywords and abstract were taken as summary models, since they are supposed to be manually produced by some of the authors as a summary of the content of the presented work.

5.1.4 Evaluation method

To automatically establish the quality of the summarizers we decided to use the ROUGE package measures presented in Section 2.3. As the output of an ASR is not necessarily grammatical or readable, the fact that ROUGE does not capture this features of the summary is not a shortcoming in this case.

Since one of the model summaries to be compared with automatic summaries is a list of keywords without any particular order, we decided to use the ROUGE-1 measure of unigram overlap to evaluate the content overlap between summaries. Moreover, as shown in Lin and Hovy (2003), the 1-gram measure between summary pairs correlates well with human evaluations.

The following ROUGE parameters were set:

- 95 per cent confidence interval is used.
- The stemming option is set.
- Stop words are not included in the calculations.
- A length limit is imposed.

5.2 Original textual FEMsum instantiations

The SDS FEMsum instantiations, *LCsum* and *MLsum*, presented in Chapter 3, were initially designed to summarize textual news. To analyze if the same strategies can be adopted when changing the type of document to be summarized, the original FEMsum instantiations, without adaptation, were evaluated with manual transcripts (corpus in Section 5.1.1) and automatic transcripts (corpus in Section 5.1.2).

The evaluation of these instantiations with manual transcripts is intended to analyze the performance of FEMsum when changing domain and genre. Manual transcripts allow us to carry out a first study avoiding the problems specific of ASR. We expect that the discourse structure and the vocabulary used in journalistic documents will not be the same as the one used in spontaneous speech transcripts, so this evaluation will help to detect specific features in these different input documents.

The work carried out with manual transcripts is presented in Section 5.2.1, the one with automatic transcripts in Section 5.2.2 and Section 7.3 analyzes the results obtained using original FEMsum instantiations.

5.2.1 Summarizing manual transcripts

This section presents the performance of the original FEMsum SDS instantiations when summarizing manual transcripts. The corpus presented in Section 5.1.1 has been used as test corpus.

The evaluation is carried out at two levels. On the one hand, the impact of producing summaries of different lengths (10, 30 and 70) is analyzed. On the other hand, we compare results using spontaneous speech, extract-based summary models vs. using well-written, abstract-based summary models.

In order to take advantage of the information in paper-based models, we have studied the relation between transcript, extract-based manual summaries (H1: human1 and H2: human2) and paper, abstract-based summaries (T: title, K: list of keywords and A: abstract). The assumption is that paper summaries (the title, the list of keywords and the abstract of the paper) are good summary models for the corresponding talk.

However, it is known that an extract-based summarizer obtains better results when evaluated against extract-based models. For that reason, the two sets of summaries produced by the assessors (Human 1 and Human 2) have been evaluated as automatic summaries, using the paper information as gold corpus. The score obtained by the assessor ideal summaries can be considered as a kind of upper bound to be achieved by an extract-based summarizer. See the two first rows of Table 5.2 for 10-word summaries, Table 5.3 for 30-word summaries and Table 5.4 for 70-word summaries.

For 10-word summaries, two paper models were used: the title (T) and the list of keywords (K). For 30-word summaries a single paper model was used: the concatenation of the title and the list of keywords (T-K). Finally, for 70-word summaries the model was the abstract (A).

In addition, to determine the quality of the paper abstract-based summaries, their performance has been evaluated considering as reference the set of summaries produced by the two assessors. The third and fourth rows in Table 5.2 and the third row in Table 5.3 and Table 5.4 reflect the scores for paper-based summaries.

For each summary length, each FEMsum approach, *MLsum* and *LCsum*, has been evaluated taking into account three different summary models. For 10-word summaries, *MLsum* has been evaluated with the sentence compression module, *MLsum(HL)*. In general, as *LCsum* and *MLsum*

	paper	**assessor**	**paper+assessor**
	T+K	H1+H2	T+K+H1+H2
Human 1	**43**	–	–
Human 2	35	–	–
Title	–	39	–
Keywords	–	26	–
MLsum(HL)	10	13	12
MLsum	14	20	18
LCsum	*25*	*35*	*31*

Table 5.2: ROUGE-1 scores taking as reference: 2 extract-based human summaries, 2 abstract-based author paper summaries (T: title, K:list of keywords) or all of them as 10-word summary models. Types of evaluated summaries: human ideal automatic systems (**upper bound**), paper abstract based, and FEMsum (*best approach*).

	paper	**assessor**	**paper+assessor**
	T-K	H1+H2	T-K+H1+H2
Human 1	**48**	–	–
Human 2	43	–	–
Title-Keywords	–	30	–
MLsum	26	26	26
LCsum	*33*	*30*	*30*

Table 5.3: ROUGE-1 scores taking as reference: 2 extract-based human summaries, 1 abstract-based author paper summaries (T-K: title and list of keywords) or all of them as 30-word summary models. Types of evaluated summaries: human ideal automatic systems (**upper bound**), paper abstract based, and FEMsum (*best approach*).

extract complete TUs, the ROUGE scores are computed considering the N (10, 30 and 70) first words of the produced summary.

The first column of each table presents the results obtained when using abstract-based paper summaries as models. In the second column, extract-based manual summaries are the models. The last column presents the performance of the system evaluated by comparison to both types of models.

Analysis of the results

Observing the three tables it can be concluded that, in the three sets of experiments, *LCsum* is always better than *MLsum*. In addition, we can see that:

	paper	**assessor**	**paper+assessor**
	A	H1+H2	A+H1+H2
Human 1	**28**	–	–
Human 2	24	–	–
Abstract	–	9	–
MLsum	17	25	22
LCsum	*19*	*27*	*24*

Table 5.4: ROUGE-1 scores taking as reference: 2 extract-based human summaries, 1 abstract-based author paper summaries (A: abstract) or all of them as 70-word summary models. Types of evaluated summaries: human ideal automatic systems (**upper bound**), paper abstract based, and FEMsum(*best approach*).

- Worst results are obtained when automatic summaries are compared to abstract-based models only (see first column of each Table: T, K in Table 5.2; T-K in Table 5.3; and A in Table 5.4).
- Best results are obtained when automatic summaries are compared to extract-based summaries only (H1 + H2, see second column of each Table).
- The scores obtained by artificial paper systems (Title, Keywords, Title-keywords) are comparable to the ones obtained by *LCsum*: in Table 5.2, when compared to human extract-based summaries, *LCsum* obtains *35* for 10-word summaries, while Title obtains 39 and Keywords obtain 26. In Table 5.2, for 30-word summaries, the concatenation of Title and Keyword obtain 30, the same score as *LCsum*. But, the score obtained by the Abstract of the paper is not competitive at all, it scores 9 when compared to the human extract-based summaries, in contrast to the *27* obtained by *LCsum* (Table 5.4).

It has to be said that the obtained scores are also affected by the fact that two of the transcripts have no paper and in some of the papers there is no list of keywords. This means that some artificial paper systems produce empty summaries.

LCsum summary analysis

When analyzing the corpus it has been observed that manually transcribed TUs are not always as good as expected. Figure 5.3 shows an example of a TU of 169 words selected as a summary. It is also a good example of the vocabulary and the kind of discourse structure found in transcripts, often highly speaker dependent. In this case, we find *of course*, a discourse marker used quite often in spontaneous speech, especially by this speaker, but not much in journalistic documents. The POS tagger has been trained with written documents, with few samples of this occurrence, for that reason the word “course” has been labeled as a common name, being considered as a strong lexical chain member. This shows how misleading lexical chains may be found and taken as reference to select the summary TU.

> Thank you as you already heard this is the title of the paper and **it**’s cooperation between Lund University and KTH in Stockholm and **it**’s a small project actually in within the Swedish language technology program **it**’s only half man year on each place but what we tried to do is *of course* to look into this prosody of Swedish with different goals in mind *of course* to attain n new knowledge about phrasing in Swedish and what we are looking a at **it** from different perspectives both the phonology and how the phrasing is actually implemented in looking at different acoustic signals that are used for implementing phrasing *of course* **it** might be obvious to most of you but talking about phrase boundaries you think about boundaries but *of course* **it**’s not necessarily boundaries **it** could easily also be coherent signaling signaling the chance not rather than the boundaries and th that will be obvious that **it**’s exploited quite a lot in in Swedish prosody as we see.

Figure 5.3: Example of a TU of 169 words.

The fact that pronouns are highly used in spontaneous speech and the reference is not always clear (see the use of **it** in Figure 5.3) is another important feature of spontaneous speech to be taken into account.

Another relevant aspect, depicted in Figure 5.4, is that, in general, the process of NERC is not adequate for the new domain. An example of OTHERS, used when the NERC system is not able to classify with some of the predetermined classes, is presented. Moreover, presented systems, algorithms, methods (e.g. CELP) are often classified as ORGANIZATION.

```
NE_class (lemma)  [TU_id(lema)]
OTHERS: (Dutch_English_German) 29(Dutch) 29(German) 33(French_Dutch) 33(German)
 36(Dutch) 36(French) 36(German) 36(German) 36(French_German) 36(Dutch) 50(Dutch)
 58(Dutch) 58(Dutch) 70(Dutch) 82(English) 83(Dutch) 83(French) 83(German) 83(Dutch)
 83(Dutch) 86(English) 86(French) 86(Germanic)
ORGANIZATION: 2(TFI) 4(TFI) 8(TFI) 10(TFI) 12(TFI) 13(TFI) 19(TFI) 25(TFI) 28(TFI)
 28(TFI) 29(TFI) 30(TFI) 31(TFI) 36(TFI) 37(TFI) 37(TFI) 60(TFI) 61(TFI) 62(TFI) 63(TFI)
ORGANIZATION: 3(CELP) 18(CELP) 27(CELP) 37(CELP)
```

Figure 5.4: NE lexical chain examples.

Figure 5.4 presents a collateral effect of the inadequate NERC process. Dutch, English and

German are considered to refer to the same proper name as Dutch_English_German. For that reason, all of them appear as members of the same chain. The problem comes from the fact that Dutch_English_German has been wrongly labeled as a proper name. A simple algorithm is used by the *LEXICAL CHAINER* to detect that two proper names are referring to the same entity. This algorithm takes into account the existence of a partial match to consider the existence of a shorter way to denote the same entity (e.g. "Romano_Prodi" and "Prodi" or "R._Prodi").

5.2.2 Summarizing automatic transcripts

The aim of this study was to orient the adaptation of the system when, instead of having well-written text as input, the input is automatically transcribed spontaneous speech. The first important difference is that the new input is ungrammatical and has no capitalization or punctuation marks.

Besides the fact that the performance of ASRs is still far from perfect, one of the most challenging problems in changing the media of input documents is to determine how to scale the new kind of documents. For the original *LCsum* and *MLsum*, sentences or syntactic segments were good units of meaning. Due to the lack of punctuation and grammaticality of the output of ASRs, determining what a TU is becomes a first new challenge.

The rest of the section as structured as follows: First, the segmentation method used for the evaluation of the original FEMsum approaches is presented. Then, we compare the performance of FEMsum when summarizing the corpus presented in Section 5.1.2, automatically segmented, manually segmented and the corresponding manual transcripts. Finally, in order to find the best way to adapt FEMsum, the properties of some automatic transcripts are contrasted with the news used in DUC.

TU segmentation

Several methods have been proposed to deal with the segmentation of dialogue transcripts into units. In general, they are based on the use of statistical or ML techniques for automatic learning, such as those in Lavie et al. (1994) (segmentation of dialogues into Semantic Dialogue Units) or Stolcke and Shriberg (1996) and Zechner (2001) (segmentation of dialogues into sentences or partial sentences with or without syntactic structure). However, the concept of segment in a dialogue is quite different than in a monolog, which is our challenge.

In the framework of the CHIL project we proposed and evaluated two segmentation methods: one heuristic-based and the other ML-based.

The first one considers a segment as a sequence of approximately N words in length, with the restriction that it should not split a syntactic chunk. When this restriction is not satisfied, the whole chunk is added to the previous segment and the new one starts after it. Chunks in documents are detected using the Yamcha chunker (Kudo and Matsumoto 2001). The objective is to avoid very long sentences, which are not useful for Text Summarization (TS).

For the second segmentation method, several experiments using SVMs were carried out. The proposed method works as follows: a sliding window of a fixed size is used and a feature vector around each word in the transcription is generated. The studied features were: word forms and lemmas, as well as POS and chunking information. Using the learned model, the feature vector is classified into one of two classes to tell whether the word is the beginning of a new segment (class B) or it is inside the present one (class I). This process is done for each word in the transcript in a sequential way. After that, a post-process is executed in order to obtain the segmented transcript from the labeled sequence of words.

The ML-based TU segmentation method and the obtained results were deeply analyzed in (Fuentes et al. 005a). However, the textual segments obtained with this method were much too long to be used in TS tasks. For that reason, for the experiments reported in the rest of the section, transcripts are segmented using the heuristic-based approach.

Automatic vs. Manual segmentation

Considering the properties of the targeted new input, automatic transcripts, we have carried out a set of experiments. As the architecture of the system presupposes some kind of natural text segmentation in order to classify each segment as being part of the summary or not, we decided to contrast the manually versus the automatically punctuated input.

To contrast the performance when using manual transcripts or automatic transcripts we have reproduced the evaluation of the system with manual transcripts but summarizing only the 29 documents from the corpus presented in Section 5.1.2. Table 5.5 shows the performance of *MLsum* and *LCsum* when summarizing manual transcripts, manually segmented automatic transcripts and automatically segmented automatic transcripts. For the evaluation, only extract-based human summaries are used as models to compute ROUGE-1 scores.

As it was expected, both approaches perform better when summarizing manual transcripts, and worst when summarizing automatically segmented automatic transcripts. This shows that the summarization process is highly sensitive to noise in the input. Moreover, as it was observed in the first evaluation with manual transcripts in Section 5.2.1, *LCsum* always performs better than *MLsum*.

As the TU classifier instantiated by *MLsum* has learned rules based on the training corpus, differences in corpus properties can easily affect the performance of this approach. To get a better insight of why *MLsum* performs worse and to help in adapting FEMsum approaches, a detailed corpus analysis is presented in the following section.

Approach	10-word	30-word	70-word
manual transcripts manually segmented			
MLsum	24	27	24
LCsum	35	28	27
automatic transcripts manually segmented			
MLsum	13	21	22
LCsum	29	24	25
automatic transcripts automatically segmented			
MLsum	8	8	20
LCsum	22	19	20

Table 5.5: Performance of the FEMsum approaches producing 10,30,70-word summaries for 29 transcripts.

Corpus comparative study

This section compares some characteristics of the 8 TED corpus presentations transcribed by JANUS (see Section 5.1.2), with the 147 DUC documents used for training (see Section 3.6.2).

Table 5.6 shows that the documents have important differences in the number of TUs per document (61 vs 93.1), the number of words per TU (27.8 vs 20.6) and in the behavior of lexical chains, although there is not a big difference on the average number of words per document.

Corpus Data	**DocLength (TU)**	**TULength (words)**	**Strong Chain**	**Strong Chain Score**
TED	61	27.8	4 [33/8doc]	20.4
DUC	93.1 [13-216]	20.6	0.6 [93/147doc]	7.8

Table 5.6: Corpus comparative study.

Another observation is that, although NEs or proper nouns are not recognized, due to the lack of capitalization and punctuation, the number of strong lexical chains that crosses on average a TU is bigger in TED transcripts than in DUC news. Moreover, lexical chain scores are higher (20.4 vs 7.8).

Table 5.7 shows 4 examples of strong lexical chains. The second column indicates the score achieved by each one of such chains. These scores are directly proportional to the number of terms related to the same concept or lexical chain member (first column) occurring in a document. We have observed that, in general, lexical chains reflect the use of frequent term repetition in spontaneous speech transcripts.

Lexical chain member	**Lexical chain score**
dialogue	13
system	29
user	15
utterance	12

Table 5.7: Example of strong lexical chains members and their score.

To finish with the analysis of features, it has to be said that the weighted cosine measure, used as a feature by *MLsum*, considers term frequency in a textual corpus of news, so it can be considered a domain dependent feature. To conclude, if we want to properly adapt *MLsum* and try to obtain better results, a new TU classifier should be trained with a set of documents that are similar to the documents in the test set.

5.2.3 Analysis of the results

The previous experiment results show that the performance of the lexical chain based approach, *LCsum*, is always better than the one obtained by the ML-based one, *MLsum*. Then, it can be concluded that the first one is less domain, style or media dependent than the second one. Moreover, the use of lexical chains seems to be a good way of detecting relevant information in spontaneous speech documents. Indeed, it has been observed that, in oral language, speakers frequently repeat important concepts. It has been confirmed that errors introduced in preprocessing of documents, both introduced by the ASR and by the text segmenter, significantly affect the performance of the approaches.

With respect to *MLsum*, in general, the main problem of adapting a ML-based approach is that dealing with a new domain requires learning a new model. In our case, the initial model for summarizing was learned using journalistic written documents. This model is not useful in the CHIL scenario. Then, a new model for the technical speech domain is required.

It must be said that the corpus used in this first set of experiments is not big enough to obtain a statistically significant evaluation. In Table 5.8 it can be observed that depending on the number of documents used in the evaluation process, the performance of the system suffers important changes. The values are almost the same when summarizing 29 and 39 documents, but important differences can be found when the number of summarized documents are 34.

The values obtained in these experiments can be used as an orientation, but either 29 or 39 are not enough documents to carry out a stable evaluation that allows to draw strong conclusions. Given the enormous cost of building a comprehensive corpus for summary evaluation, we propose to use the title and the list of keywords from the paper that is being presented as summary models, to be used as a gold standard to assess the quality of summaries of the oral presentation transcript. However, paper abstracts have to be discarded, as the comparison of

Approach	10-word	30-word	70-word
29 manual transcripts			
MLsum	24	27	24
LCsum	35	28	27
34 manual transcripts			
MLsum	22	**33**	**46**
LCsum	34	**43**	**58**
39 manual transcripts			
MLsum	20	26	25
LCsum	35	30	27

Table 5.8: Performance of the FEMsum approaches producing 10,30,70-word summaries for 29, 34 and 39 manual transcripts.

longer summaries is almost impossible, due to the important differences between texts of extracts from presentations and paper abstracts (discourse structure, vocabulary,...). This problem also affects the comparison of 10- and 30-word automatic summaries with the title and keywords, but stylistic differences are less dramatic in shorter texts.

5.3 Exploiting general textual properties

To tackle the variation in register and the noise of faulty ASRs, we propose the use of robust, domain-independent summarization approaches. Lexical chains and Discourse Marker (DM) information are exploited to detect relevant information in transcripts of monologues. In contrast to other state-of-the-art systems, acoustic information has not been considered because it is not always available.

The results obtained in the experiments presented in Section 5.2 show that exploiting textual properties is specially suitable for the problem of summarizing spontaneous speech, because in oral presentations important concepts tend to be highly repeated and discourse markers are frequently used. In addition, the success of Stokes et al. (2004) in exploiting lexical chains as text representation to obtain short summaries of broadcast news seems to indicate that *LCsum* is robust enough to deal with different registers. We present here two summarization *LCsum* adaptations, *PreSeg* and *PostSeg*, whose architectures (see Figure 5.5) share a highly portable core. This core relies on domain and register independent linguistic processes, and specific modules are added as required.

As depicted in Figure 5.5, *PreSeg* and *PostSeg* differ in the way they tackle the task of segmenting text: in the *PostSeg* approach the identification of segment boundaries is deferred until relevant content is detected, whereas in the *PreSeg* approach the text is segmented at the

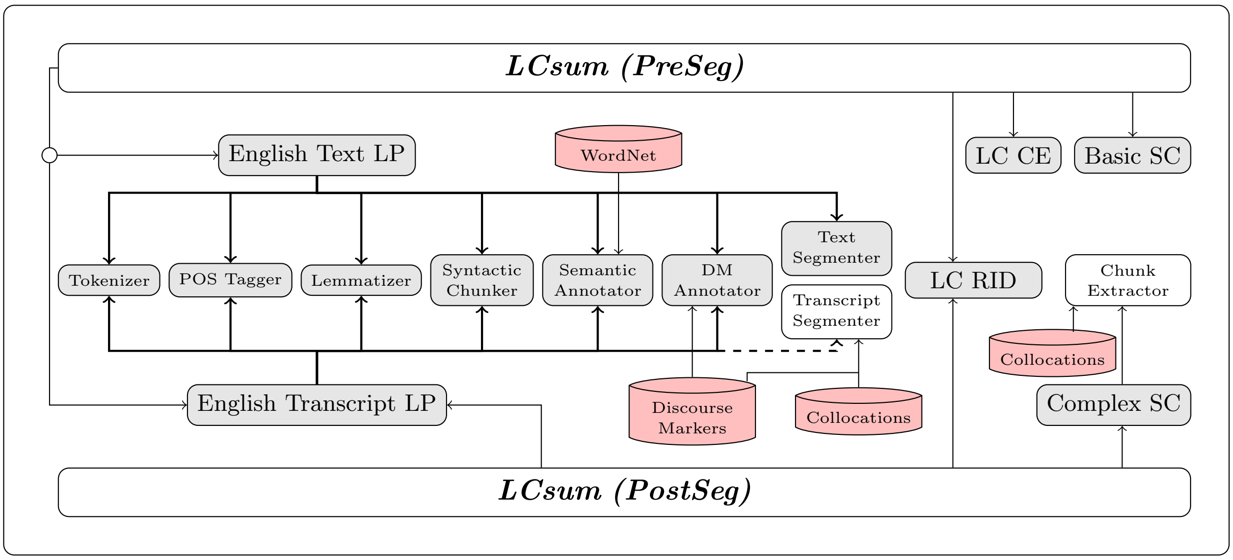


Figure 5.5: *PreSeg* and *PostSeg* modules.

linguistic preprocessing step, as detailed below.

Two external resources are required for segmentation: a list of collocations and a list of DMs[3] (Alonso (2005), Appendix A). Collocations[4] are extracted automatically from a written corpus applying a χ^2 hypothesis test (as proposed in Manning and Schütze (1999)) to all n-grams of up to a certain size. In the *PostSeg* approach only collocations filtered by syntactic patterns are considered (a subset of those in Arranz et al. (2005), see Table 5.9).

definite−{JJ}_clause−grammar{NN}
non-iterative{JJ}_matrix−inversion{NN}
bottle{NN}_necks{NNS}
gradation{NN}_of{IN}_spontaneity{NN}
bionic{JJ}_wizard{NN}
spaces{NNS}_of{IN}_perceptual{JJ}_distinction{NN}
correlates{VBZ}_of{IN}_perceptually{RB}

Table 5.9: Example of collocations filtered by syntactic patterns.

As can be seen in Table 5.10, each DM has associated a value of relevance: *dm* neutral (purely formal relevance), *nuc* signals relevance and *sat* signals lack of relevance. Furthermore, syntactic classes are also distinguished: *r* for DMs that dominate words to their right or *b* for

[3] `http://russell.famaf.unc.edu.ar/~laura/shallowdisc4summ/discmar/`

[4] sequences of words that are likely to co-ocur.

DMs dominating words at both sides. Every DM is associated to a score of rhetorical relevance (2, 1 or -1). This score is assigned to all words dominated by the DM (following or preceding the marker, depending on its syntactic type).

discourse marker	**relevance**	**syntactic**
because	dm	b
although	sat	r
first_of_all	dm	r
however	nuc	r
in_case	sat	b
in_order_to	nuc	b

Table 5.10: Example of discourse markers, their relevance and syntactic class.

As depicted in Figure 5.5, the main tasks performed by the adapted *LCsum* approaches are the following:

1. **LexicoMorpho LP**: As described in Section 3.4, in this step generic NLP processing tasks are performed over an input text (tokenizing, tagging, lemmatizing, syntactic chunking and semantic labeling, identification of DMs and collocations) [5].

   To extract the collocation list, in this case domain-specific, we have used 429 EuroSpeech'93 conference papers, trying to obtain domain-specific collocations.

   In the *PreSeg* approach, the input text is segmented at this point, as detailed in the following item. In the case of *PostSeg*, the decision is deferred to the SC component.

   1a. **PreSeg Transcript Segmenter**: Segments of n words are identified in the input text. A size of 20 words per segment has been established, similar to the average sentence length in the transcript corpus (20.81), with the restriction that segment boundaries cannot be placed

   – before a coordinating conjunction,
   – after a conjunction, a preposition or a determiner,
   – so that they split a syntactic chunk or a collocation,
   – so that they leave a DM in the last m positions of the segment (where m is related to the scope size of the DMs). Since oral input is ill-formed and the scope of the DMs cannot be calculated in terms of linguistic units, the scope has been stipulated as a 5-word window following it or a 10-word window around it, depending on the syntactic type of the marker.

[5] In most cases, no adaption has been done and the preprocessors are based on written text models.

When a DM is found and we need less than m words to complete a segment, a boundary is placed immediately before it, yielding a segment shorter than n words. As for the rest of restrictions, words are added to the segment, until a suitable splitting point is found, possibly producing segments longer than n.

2. **Lexical Chain based RID**: As described in Section 3.6.1, in order to recover most of the information in the text, we consider medium and light chains, instead of only strong ones. This measure is aimed to minimize the bias of strong chains in spontaneous speech: they tend to provide a misrepresentation of the information in a text because the distribution of the frequency of words is rather skewed, and only few strong chains are found. In all the experiments, lexical chains were computed taking into account only common nouns as candidate chain members. Due to the lack of punctuation and capitalization, usual textual methods to detect NEs are useless. Only repetitions and synonyms have been considered, disregarding other Wordnet relations. In this domain, very general words (e.g., *speech or speed*) are used with a very specific sense, but since no domain-specific word-sense disambiguation is performed, considering all their WN relations would yield an inadequate representation of the text.

3. **Lexical Chain based CE**: In the *PreSeg* approach, *Heuristic 2*, ranking highest the TU crossed by the maximum of *Strong* chains, is set.

4. **Complex SC (Chunk Extractor)**: The implementation of this step is not the same in both approaches. While *PreSeg* uses the Basic SC, *PostSeg* determines at this point the segment boundaries by using the Chunk Extractor described in Section 3.8.1. In the Chunk Extractor, 20 is the defined chunk size.

The performance of the *PreSeg* and *PostSeg* has been evaluated at two levels, producing summaries of different lengths (10 or 30 words) and having manual or automatic transcripts as input. As in the first set of experiments, in order to establish an upper bound for extract-based summarizers, the summaries produced by assessors were evaluated as if they were the output of two ideal systems (Human 1 and Human 2).

In order to have a fair comparison between manual and automatic transcripts, capitalization was removed from the former except in the *ManSeg* where the original *LCsum* summarizes the manual transcripts manually segmented.

Two kinds of baseline summaries were also evaluated:

- NFirst, the first n (10 or 30) words of each transcript.
- NFreq, the n-long segment that maximizes the number of frequent words contained in it.

Table 5.11 shows the results of the experiments when comparing extract-based summaries from scientific presentation transcripts against three sets of models: the title and keywords of the corresponding paper (*paper*), both manual summary models (*assessor*) and all of them (*paper + assessor*).

Summary Length	**10-word**			**30-word**		
Summary references	**paper**	**assessor**	**paper+assessor**	**paper**	**assessor**	**paper+assessor**
	T+K	H1+H2	T+ K+H1+H2	T+K	H1+H2	T+K+H1+H2
	manual transcriptions (37 oral presentations)					
Human 1	44.87	–	–	55.26	–	–
Human 2	37.38	–	–	50.12	–	–
NFreq	16.66	18.18	17.85	23.28	21.84	22.54
NFirst	19.88	26.67	24.40	34.76	20.92	24.92
ManSeg	26.39	34.59	31.06	41.07	29.90	32.75
PreSeg	24.99	32.11	29.04	38.67	33.66	35.25
PostSeg	26.16	31.65	29.20	39.61	32.04	34.13
	automatic transcriptions (93 oral presentations)					
NFreq	18.61	–	–	18.46	–	–
NFirst	21.83	–	–	28.70	–	–
LCsum	24.39	–	–	31.90	–	–
PreSeg	25.76	–	–	38.55	–	–
PostSeg	29.70	–	–	49.98	–	–

Table 5.11: Upper bound (Human ideal automatic systems) and system performances when summarizing manual or automatic transcripts. ROUGE unigram overlap measure have been computed when taking, extract-based human summaries (H1:human1 and H2:human2), abstract-based author paper summaries (T:title and K:list of keywords) or all of them as model summaries.

When summarizing automatic transcripts, besides the baseline and the two adaptations of *LCsum*, the performance of the original *LCsum* approach, evaluated in Section 5.2.2, is also analyzed.

5.3.1 Analysis of the results

As could be expected, Table 5.11 shows that recovering the content of model summaries in 10 words is more difficult than in 30 words.

It is noteworthy that in 10-word summaries automatic summaries are more similar to assessor summaries than to the title and keywords of the paper. However, this tendency is reversed for 30-word summaries.

A t-student test with 95% confidence interval was applied to determine whether differences between the summaries produced by the different automatic summarizers (detailed in Table 5.11) were significant. On the one hand, when summarizing manual transcripts, assessors always perform significantly better than any automatic system, as compared to paper-based summaries. For 10-word summaries, Human1 summarizes better than Human2. Differences between *ManSeg*, *PostSeg* and *PreSeg* are never significant, but the *NFreq* baseline is clearly the worst automatic system, and *Nfirst* is worse than *ManSeg*. On the other hand, when having as input automatic transcripts, *PreSeg* and *PostSeg* perform significantly better than any baseline.

It is interesting to note that the quality of summaries does not drop when automatic, instead of manual, transcripts are used. This shows that the proposed techniques are robust enough to exploit salient textual features that allow to identify and delimit the most relevant parts of a presentation.

To analyze the obtained scores and determine the agreement between the different models, we have studied the correlation between them. Taking as a reference one model at a time, we have computed the ROUGE-1 measures for each kind of automatic summary. After that, the level of agreement between pairs of models has been measured comparing the values given to each summary according to the models. The Pearson correlation coefficient has been used to quantify the agreement between pairs of models, reaching a 99% confidence level in all cases, which indicates that different models agree significantly. Then, it can be concluded that the ranking obtained by the evaluated approaches is not dependent on the summary model taken as reference. Table 5.12 displays correlation values for each pair of models.

Manual-10	Title	Human1	Human2
KeyWords	0.567	0.529	0.495
Title		0.646	0.633
Human1			0.744

Manual-30	Title	Human1	Human2
KeyWords	0.457	0.493	0.397
Title		0.565	0.340
Human1			0.608

Automatic-10	Title
KeyWords	0.661

Automatic-30	Title
KeyWords	0.679

Table 5.12: Pearson correlation coefficients between the different models in each evaluation set.

For 10-word summaries from manual transcripts, one of the assessors correlates slightly better with paper models (KeyWords and Title) (0.53 and 0.65, respectively) than the other assessor (0.50 and 0.63). The same tendency is accentuated in 30-word summaries: 0.49 and 0.56 for the first judge, 0.40 and 0.34 for the second. This explains the fact that, when compared to paper-based summaries (see Table 5.11), Human2 obtains lower performance than Human1

(37.38, 50.12 *vs.* 44.87, 55.26, respectively). The style of assessor extracts seems closer to the titles of the papers than to the list of keywords, although differences between different kinds of paper-based summaries and assessor extracts are smaller in 30-word summaries. In any case, it is interesting to note that the correlation between Human1 and Human2 is the highest of all in both summary sizes (0.74 for 10-word, 0.61 for 30-word). The correlation between the keywords (K) and title (T) models is also significant, especially when evaluating automatic transcripts (0.66 and 0.68).

5.4 Conclusions

We have manually built an evaluation corpus in the framework of CHIL project, to compare the performance of the different FEMsum approaches to summarize automatically transcribed oral presentations.

The values obtained using this evaluation corpus can be used as an orientation, but it does not contain enough documents to carry out a confident evaluation. Given the enormous cost of building a comprehensive corpus for summary evaluation, we propose to use the title and the list of keywords from the paper presented in the oral presentation as a gold standard summary.

Correlation between models has been studied, showing that the title and the list of keywords of the paper are good summary models to evaluate scientific presentations summarizers, although human made summaries are more stable. In contrast, the Abstract of the paper can not be used as a model due to style differences between well-written text and spontaneous speech.

To summarize transcripts from oral scientific presentations, several NLP processing tasks need to be adaptated: NER, NEC and other NLP tools that have been trained in other domains. In the reported experiments, the tools used are based on written text models. We have focused our efforts on the adaptation of the segmentation in textual units for transcripts. Two different approaches to segmentation have been implemented: *PreSeg* and *PostSeg*. Both approaches are adaptations of the original *LCsum*.

We have found that automatic summaries from a presentation are reasonably similar (contentwise) to human ones, and also to the title and the keywords of the corresponding paper. *PreSeg* and *PostSeg* perform significantly better than two dummy baselines. The first baseline is the first fragment of the talk, where we expect the speaker to synthesize the aim of the talk. The second baseline consists in selecting the speech segment that maximizes the total frequency score, summing up the frequencies of all the words in it. In addition, the quality of summaries does not drop when automatic, instead of manual, transcripts are used.

The *MLsum* system presented in Section 4.2 has not been adapted due to the fact that the rules of the TU classifier were learned from a totally different domain and performed rather poorly on transcripts of spontaneous speech.