Summarizing a multimodal set of documents in a Smart Room

Maria Fuentes, Horacio Rodríguez, Jordi Turmo

TALP Research Center, Technical University of Catalonia (UPC)
mfuentes@lsi.upc.edu, horacio@lsi.upc.edu, turmo@lsi.upc.edu

Abstract

This article reports an intrinsic automatic summarization evaluation in the scientific lecture domain. The lecture takes place in a Smart Room that has access to different types of documents produced from different media. An evaluation framework is presented to analyze the performance of systems producing summaries answering a user need. Several ROUGE metrics are used and a manual content responsiveness evaluation was carried out in order to analyze the performance of the evaluated approaches. Various multilingual summarization approaches are analyzed showing that the use of different types of documents outperforms the use of transcripts. In fact, not using any part of the spontaneous speech transcription in the summary improves the performance of automatic summaries. Moreover, the use of semantic information represented in the different textual documents coming from different media helps to improve summary quality.
Keywords: Summarization, Evaluation Methodologies, Multimedia Document Processing

1. Introduction

In the last years the research on multimodal Human-Computer Interaction (HCI) is evolving from basic interpretations of the signals to richer semantic information. This research area combines information from several modalities: speech, vision, language, text. Only collaborative undertakings could address the full complexity of human interaction.

The analysis of human interaction has attracted significant interest in the literature, being interactive lectures and meetings the central theme of analysis in several international projects: The AMI (Augmented Multiparty Interaction) and the AMIDA (Augmented Multiparty Interaction with Distance Access) European Integrated projects, whose goal is the support and analysis of multi-modal interactions between people in meetings with small number of participants. The US CALO project (Cognitive Assistant that Learns and Organizes). Also in a meeting scenario, CALO has developed a meeting assistant focused on advanced analysis of spoken meeting recordings, along with related documents, including emails. The CHIL European project (Computers in the Human Interaction Loop) has explored the use of computers to enhance human communication in smart environments, especially within lectures and post-lecture discussions. The research reported in this article concerns an intrinsic automatic summarization evaluation in the CHIL scientific lecture domain.

Automatic Summarization (AS) consists in "to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs" (Mani and Maybury, 1999). AS strongly depends not only on the properties of the document, but also on the user needs (e.g., size of the summary, output media, content related to a query).

Studies such as the one carried out by (Shriberg, 2005) show that oral communication is harder to process than written text. In addition, large amount of effort is required to train Automatic Speech Recognizers (ASR) models. It has to be taken into account the fact that these models are language and domain dependent. For those reasons, we propose using a multidocument summarization (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.

The structure of this paper is as follows: next section gives an overview of related research in automatic summarization and summarization evaluation. Section 3 describes the CHIL lecture corpus used for these experiments. Section 4 presents configuration details of the evaluated summarization approaches. Section 5 analyzes the results. Section 6 concludes the paper.

2. Related Work

Mostly, automatic summarization today is based on sentence-extraction paradigm for English documents, in both single document (SDS) and multidocument summarization (MDS). Knowledge-rich approaches either extend basic methods by the incorporation of sophisticated, yet general lexical resources such as WordNet (Fellbaum, 2008), or apply discourse organization theories in generic contexts, or bring domain knowledge to the summarization enterprise. Lexical cohesion has long been considered a key component in assessing content relevance in text summarization and computation of lexical chains (Barzilay, 1997) has been used as text interpretation mechanisms for selecting key sentences. There has been a growing interest on applying graph-based representations, for instance (Mihalcea and Tarau, 2005) uses a PageRank algorithm to summarize Single and Multiple documents.

With respect to spontaneous speech, less effort has been devoted. Most of the research focuses on broadcast news, usually generated by reading aloud written text.

typically read aloud from a written text. Current work on oral presentations tends to deal with a single document, a speech transcript, (Hirohata et al., 2005), (Fuentes et al., 2005), (Chatain et al., 2006), speech signal of lectures (Hori et al., 2003), or a combination of them (He et al., 2000). However, (Zhu et al., 2009) deals with summarizing multiple spontaneous spoken documents from untranscribed au-

dio. For meeting summarization, (Xie et al., 2008) mainly used maximum marginal relevance (MMR). Given a query that encodes user's information needs, MMR iteratively selects the textual units most relevant while trying to avoid redundancy.
The evaluation has also become a critical issue for very complex Natural Language Processing (NLP) applications, such as AS. According to (Sparck-Jones and Galliers, 1996), two major types of NLP evaluation can be differentiated: *intrinsic* and *extrinsic* evaluations. Both can be applied to evaluate systems. The former directly measures the quality of the created summary by evaluating some intrinsic properties, such as coverage, responsiveness or readability. The later indirectly measures summary performance in a task dependant on the quality of the summary. Usually, extrinsic methods are used to compare the accuracy of different systems without assigning absolute scores to them. TIPSTER SUMMAC and the Japanese[1] NTCIR are good examples of AS extrinsic evaluations. In contrast, the ones carried out in DUC and TAC contests are good examples of AS intrinsic evaluations.
Evaluating summaries, either manually or automatically, is a hard task. The main difficulty in evaluation comes from the impossibility of building a fair gold standard against which the results of the system we wish to evaluate can be compared. This difficulty is due to the very low agreement among human evaluators when faced with interpretative decisions. This lack of agreement comes, in turn, from the difficulty of defining the set of measurable properties that contribute to the quality of summaries. Usually, different properties are taken into account and it is a difficult task to select the most appropriate ones and to combine them properly. (Mani, 2001) provides a clear picture of summary evaluation, both by human judges and by automated metrics. DUC conferences adopted the ROUGE package for automatic content-based evaluation (Lin and Hovy, 2003). ROUGE includes a series of recall measures based on n-gram co-occurrence statistics between a peer summary and a set of model summaries.

3. Evaluation framework

The approaches studied in this article have to answer a query by summarizing documents of different natures. In this task, we focused on the 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.

The Smart 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.
ELDA, in charge of the creation of the CHIL test corpora, selected 10 seminars from the ones recorded at Karlsruhe University, at the ISL. Table 1 presents an example of the ISL topics.

Seminar ID	**Topic**
20031111	Robustness through articulatory features
20041112_A	Speech translation
20041123_A	Grapheme based speech recognition
20041123_E	ISL meeting transcription system
20050112	Blind segment of acoustic signal

Table 1: Example of ISL seminar topics.

The goal is to analyze the performance of several summarization approaches when dealing with the task of 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 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).
Different sorts of English documents (4 on average) were collected for each of these technical seminars. For instance, besides the manual transcription of the seminar, additional documents from cited 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, see Table 2.

TOPIC	Grapheme based speech recognition
Query 1	Multilingual grapheme based speech recognition + poly-grapheme clustering
Query 2	Pronunciation dictionary + CART + classification and regression trees

Table 2: Example of queries for an 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. Figure 1 (at left) presents three assessors models produced for the *Speech Translation* topic, with the query: *Statistical Machine Translation + Noisy Channel Paradigm* (top part of the Figure 1).
The manually created summaries were used as models for applying several ROUGE metrics.
In particular: ROUGE-n computes an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-L computes the longest common subsequence. ROUGE-W introduces a weighting factor of 1.2 to better score contiguous common subsequences. And

[1] http://research.nii.ac.jp/ntcir/index-en.html

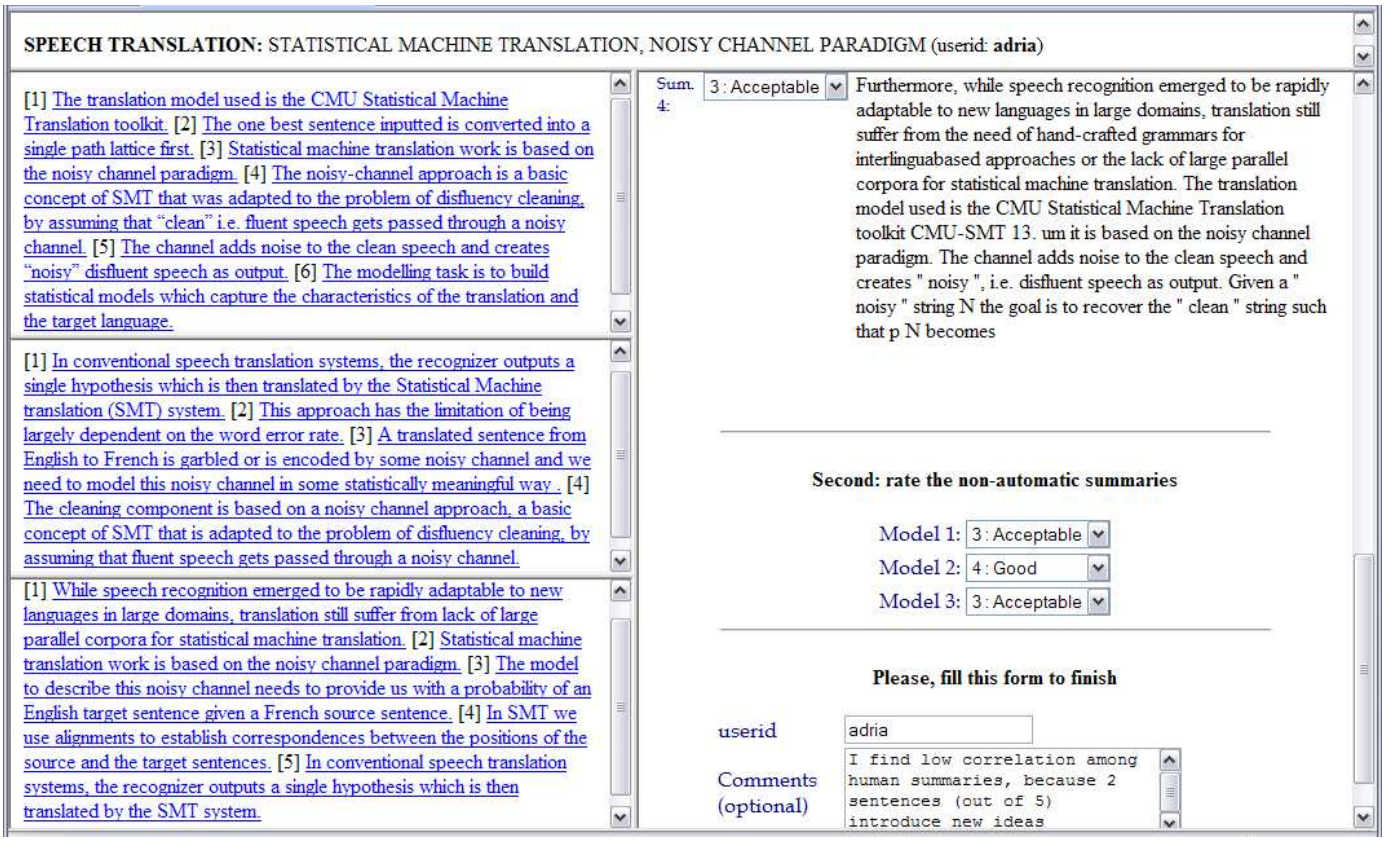

Figure 1: CHIL manual evaluation web interface.

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 1 shows the interface used by the UPC assessors in the evaluation process. 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 1). Human evaluators were also asked to score the quality of the human summaries taking into account the other two human models as reference.

4. Evaluated approaches

In the reported experiments we evaluated 5 different summarization approaches, several of them integrated in FEMsum (Fuentes, 2008). FEMsum is a flexible and highly modularized Multitask Summarization architecture, on which parameterizable Summarizers suitable to specific needs can be built. The architecture is divided in four main components: Relevant Information Detector (RID), Content Extractor (CE) and Summary Composer (SC). In addition there is a language dependent Linguistic Processor (LP), and a Query Processor (QP) component. Not all the components are needed in all FEMsum instantiations.

To allow the maximum flexibility, the main restriction of the proposed architecture is the input and the output expected by each component. The LP component enriches the original text (documents to be summarized or the user information need) with linguistic information. This component consists of a pipeline of language dependent linguistic tools. Different LP instantiations can be used depending on the requirements of the approach, as well as the language, media, genre or domain of the documents to be summarized. If a purely lexical FEMsum approach is instantiated, LP can be reduced to a segmentation task: splitting the input document into textual units, each composed by a sequence of words[2]. However, semantic based approaches can require more sophisticated language dependent tools to enrich the text with syntactic or semantic information.

Textual Unit (TU)s are enriched with lexical (*sent*) and syntactic (*sint*) language dependent representations. For each TU, its syntactic constituent structure (including head specification) and the syntactic relations between its constituents (subject, direct and indirect object, modifiers) are obtained (see an example in Figure 2). From *sent* and *sint*, a semantic representation of the TU is produced, the environment (*env*).

Env is a semantic-network-like representation of the semantic concepts (nodes) and the semantic relations (edges) holding between the different tokens in *sent*. Concepts and relation types belong to an ontology of about 100 semantic classes (as person, city, action, magnitude, etc.), and 25

[2]It is possible that the segmentation process only requires a simple word stemming or introduces some complexity detecting MWs, terms, NEs. However, straightforward methods, usually language independent, can be applied with a rather small decrease in accuracy.

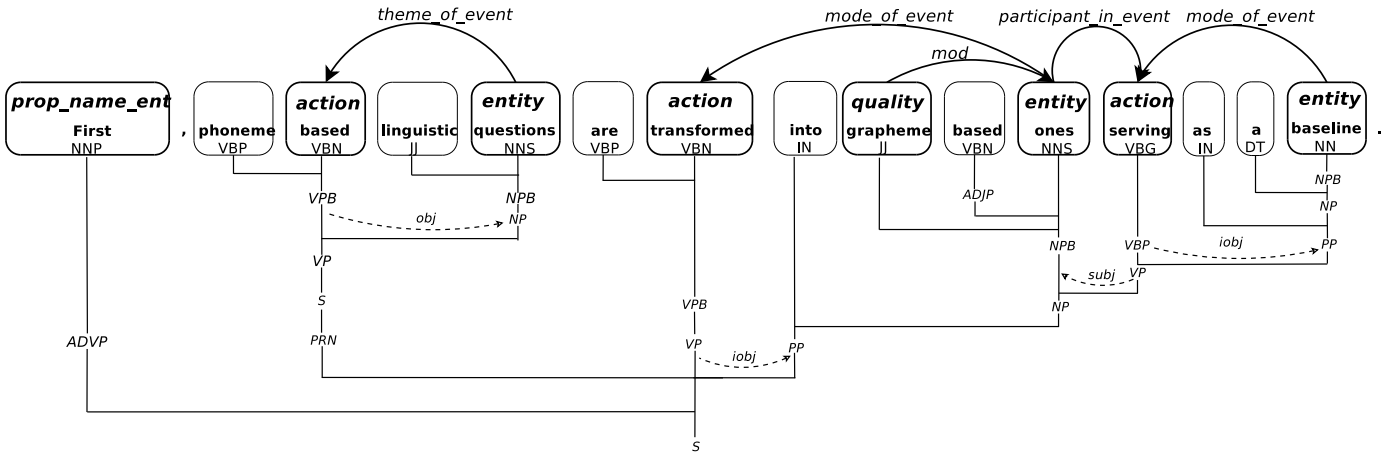

Figure 2: Example of preprocessed sentence

relations between them (mostly binary, as time_of_event, actor_of_action, location_of_event, etc.). Both classes and relations are related by taxonomic links allowing for inheritance.

The input to the RID module is the document or set of documents to be summarized. This documents can be previously enriched with some linguistic information. The original user need linguistically processed is the input of the QP component. The output of this component, expressing the user need, can be taken into account by RID to score the set of relevant TUs (segments, phrases or sentences). The output of RID is a set of TU identifiers ranked by relevance. The linguistic information and the relevance score of the TUs is the input to the CE and the SC components.

The main function of the CE component is to extract and score by relevance summary candidate TUs. This is not always instantiated, but when it is, this component allows to apply different heuristics taking into account some input aspects, such as the genre of the document (journalistic vs. scientific); some purpose aspects, such as the type of audience of the produced summaries (background vs. just-the-news); or some output aspects, such as the content of the summary (text-driven vs. query-driven).

The final text summary is the output of the SC. This component carries out the post-processing of the summary content. This post-processing takes into account the size or the format of the summary. In this component the summary TUs can be simplified, paraphrased, reordered, or removed. The following sections give a brief description of the five summarization approaches manually and automatically evaluated in the CHIL project is presented.

4.1. SEM

The RID component provides a ranked set of relevant TUs by using JIRS (Gómez et al., 2005), a Passage Retrieval (PR) software. Due to the fact that textual transcriptions from spontaneous speech are often ill-formed and they not always follow the written syntactic rules, transcription TUs have not been included in the input of the Content Extractor component. In this case, a twofold CE component scores the candidate sentences and selects the most appropriate one to be included in the summary. A similarity matrix among candidates is computed. For that purpose, the semantic representation of each candidate TU is transformed into labeled directed graph representation, where nodes are assigned to positions in the sentence and labeled with the corresponding token, and edges are assigned to predicates (a dummy node, 0, is used for representing unary predicates). Only unary and binary predicates are used. This approach on facing the summary extraction subtask follows the (Mihalcea and Tarau, 2005) line. However, instead of using only lexical measures, *SEM* (Fuentes et al., 2006) we proposed to add semantic measures to establish sentence scores.

Input: Sim be the similarity matrix,
$Candidates$ a list of candidate TUs,
Output: $Summary$ an ordered list of TUs to be included.

1. Set $Candidates$ to the list provided by the RID component.
2. Set $Summary$ to the empty list.
3. Set Sim to the matrix containing the similarity values between members from $Candidates$.
4. Compute for each candidate in $Candidates$ a score that takes into account the initial relevance score and the values in Sim. The score used is based on PageRank, as used by (Mihalcea and Tarau, 2005), but without making the distinction between input and output links.
5. Sort $Candidates$ by this score.
6. Append the highest scoring candidate (the head of the list) to the $Summary$ and remove it from $Candidates$.
7. In order to prevent content redundancy, the $S\%$ most similar TUs to the selected one (using Sim) are removed as well from $Candidates$ and the $R\%$ least scored are also removed from $Candidates$ to reduce the search space (S: 1,5%, R: 1,0%).
8. If $Candidates$ is not empty go to 4, otherwise exit.

Figure 3: Candidates Selector procedure.

Three criteria have been taken into account to do the candidate selection: Relevance (regarding the query), Density and cohesion, and Antiredundancy. The Candidate Selector procedure is described in Figure 3.

	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 3: 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 4: Responsiveness considering 3 human models when evaluating automatic summaries and 2 when evaluating human summaries.

4.2. LEXnoT and LEX

As in *SEM* a JIRS PR software is used to instantiate the RID component. Because of having a small number of documents to be summarized, all the TUs selected as relevant by the RID component are considered as a input of a simple SC component. The input of the SC are the TUs detected as relevant (this number ranges from 67 to 257 TUs - 186,75 in average). Following the JIRS ranking order, TUs are iteratively added to the final summary until reaching the desired summary length.

The main difference between *LEXnoT* and *LEX* is the input type of documents to be summarized. In *LEXnoT*, as in *SEM*, transcription have not been taken into account.

4.3. SDS

As we assume that adding textual information helps when summarizing spontaneous speech, we consider as a baseline a *SDS* based on lexical chains (Fuentes et al., 2005). This approach extracts segments of about 30 words only from the transcriptions and has been adapted to be query-driven by increasing the weight of the lexical chain members that appear in the query. WordNet is used to identify synonymy relations between words. Once chains are identified, the measure of chain strength can be classified into *Strong*, *Medium* and *Light*, depending on their score:

$$\begin{aligned} \tau &= \mu_s + 2 \cdot \sigma_s \\ Strong &= \{c \mid score_c \geq \tau\} \\ Medium &= \{c \mid \tau > score_c \geq \tau/2\} \\ Light &= \{c \mid \tau/2 > score_c\} \end{aligned}$$

where μ_s and σ_s are the average of the scores of all the chains and the corresponding standard deviation. In contrast to other approaches where lexical chains are used, if necessary, we consider *Medium* and *Light* chains in addition to the typical strong ones. That is specially useful when dealing with spontaneous speech, due to the fact that *Strong* chains tend to provide a misrepresentation of the information in a text, because the distribution of the frequency of words is rather skewed, that is due to the fact that in oral presentation important concepts are numerously repeated. For that reason only few strong chains are found. Taking into account the lexical chains found, a windows of n contiguous words (chunks) are extracted to form a summary of the targeted size. Chunks are included in the summary using a priority ranking function that tries to capture both relevance and well-formedness.

4.4. +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, we decided to evaluate the UAM-Titech06 system (Alfonseca et al., 2006). This system identifies and uses background information related to the query from the World Wide Web to produce the summaries.

5. Analysis of the results

Table 3 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.

Looking at the content responsiveness results, in Table 4, 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 final summary.

Table 5 shows the percentage of summaries classified by score. On the one hand, although SEM and +www achieve the same mean (1.8 in Table 4), Table 5 shows that the percentage of summaries considered as 'Acceptable' or 'Good' is higher in SEM (20% + 5%) than in +www (15%

	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 5: Responsiveness scores distribution by automatic system.

+ 5%). On the other hand, LEX with a lower mean score (1.775) obtains better results for acceptable summaries than SEM or +www (35%).

6. Conclusions

This paper analyzes the performance of several summarization approaches when answering a user need. The user need is expressed by a list of terms and the summary consists of a set of relevant textual fragments extracted from the document set. One approach uses only lexical features, and the other one takes into account a representation of syntactic and semantic information in order to avoid redundancy and to produce summaries with more cohesion. Both approaches use a Passage Retrieval software to detect the relevant information associated to the user's list of terms. The performance of the approaches has been studied using the summary evaluation corpus from the CHIL project. Results show that the fact of not using any part of the spontaneous speech transcription in the summary improves the performance. Moreover, the use of semantic information represented in the different textual documents coming from different media helps to improve the quality of the summaries. Although adding semantic information significantly increases the performance when dealing with written news articles(Fuentes et al., 2006), the experiments show that there is room for improvement when adding semantic information to deal with different sorts of documents from the scientific domain. This is mainly due to the fact that the language processing tools were trained on a different domain and oral presentation documents are less structured and edited than formal text.

7. Acknowledgements

This work has been partially funded by KNOW2 (TIN2009-14715-C04-04).

8. References

Alfonseca, E., Okumara, M., Guirao, J.M., and Moreno-Sandoval, A. Googling answers' models in question-focused summarisation, Proc. NAACL-HLT Workshop (DUC2006), 2006

Barzilay, R. Lexical Chains for Summarization, Ben-Gurion University of the Negev Master Thesis, 1997

Chatain, P., Whittaker, E., Mrozinski, J., and Furui, S. Class model adaptation for speech summarisation. Proc. NAACL-HLT, 2006

Fuentes, M., Gonzàlez, D., Rodríguez, H., Turmo, J., and Alonso, L. Summarizing Spontaneous Speech Using General Text Properties. Proc. Crossing Barriers in Text Summarization Research Workshop held in conjunction with RANLP, 2005.

Fuentes, M., Rodríguez, H., Turmo, J., Ferrés", D., FEMsum at DUC 2006: Semantic-based approach integrated in a Flexible Eclectic Multitask Summarizer Architecture", Proc. NAACL-HLT Workshop (DUC 2006), 2006.

Fuentes, M. A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language, Ph. D. thesis, Universitat Politècnica de Catalunya, 2008, Barcelona, Spain.

Fellbaum, C. WordNet. An Electronic Lexical Database, The MIT Press, 1998, Language, Speech, and Communication

Gómez, J.M., Montes-y-Gómez, M., Sanchos, E., Rosso, P. A Passage Retrieval System for Multilingual Question Answering, Proc. TSD, 2005.

He, L., Sanocki, E., Gupta, A., Grudin, J. Comparing presentation summaries: Slides vs. reading vs. listening. Proc. ACMCHI, 2000.

Hirohata, M., Shinnaka, Y., Iwano, K., and Furui S. Sentence extraction-based presentation summarization techniques and evaluation metrics. Proc. ICASSP2005, 2005

Hori, T., Hori, C., and Minami, Y. Speech Summarization using Weighted Finite-State Transducers. Proc. Eurospeech2003, 2003

Lin, C. and Hovy, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. Proc HLT-NAACL2003, 2003.

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

Mihalcea, R. and Tarau, P. An Algorithm for Language Independent Single and Multiple Document Summarization, Proc. International Joint Conference on Natural Language Processing (IJCNLP), Korea, October 2005

Nami, I., Automatic Summarization, John Benjamins Publishing Company, Nautral Language Processing series, 2001

Shriberg, E. Spontaneous Speech: How People Really Talk and Why Engineers Should Care. Proc. Interspeech2005, 2005.

Sparck-Jones, K., Galliers, R., Evaluating Natural Language Processing Systems: An Analysis and Review, Lecture Notes in Computer Science, Springer, 1999

Zhu, X., Penn, G., Rudzicz F. Summarizing multiple spoken documents: finding evidence from untranscribed audio. Proc. ACL and AFNLP 2009.

Xie, S. and Liu, Y., Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization, in Proc. ICASSP, Las Vegas, NV, 2008, pp. 49854988.