Knowledge-intensive automatic e-mail summarization in CARPANTA

Laura Alonso and **Irene Castellón**
GRIAL
General Linguistics Department
Universitat de Barcelona

Bernardino Casas and **Lluís Padró**
TALP
Software Department
Universitat Politècnica de Catalunya

Abstract

We present CARPANTA, an e-mail summarization system that applies a knowledge intensive approach to obtain highly coherent summaries. Robustness and portability are guaranteed by the use of general-purpose NLP tools, but it also exploits language- and domain-dependent knowledge. The system is evaluated against a corpus of human-judged summaries, reaching satisfactory levels of performance.

1 Introduction

We present CARPANTA, the e-mail summarization system within project PETRA. PETRA is related to the European project MAJORDOME - Unified Messaging System (E!-2340), whose aim is to introduce a unified messaging system that allows users to access e-mail, voice mail, and faxes from a common "in-box".

One of the lines of work developed within PETRA is the use of Natural Language Procesing (NLP) techniques for information management, namely, for text classification and summarization, as well as for information retrieval. This task includes the subgoal of text summarization, specially relevant for oral interfaces to electronic mail systems.

The summarization module within PETRA is CARPANTA. It is currently working for Spanish, but portability to other languages is guaranteed by a language-independent core.

The rest of the paper is structured as follows: first, NLP problems specific to e-mail summarization and our approach to them are described, comparing it with previous work. Section 3 presents the architecture of the system. The system is evaluated by comparison with a human-made gold standard, results can be seen in Section 4.

2 Problems of e-mail summarization

Besides the problems specific to automatic text summarization, e-mail summarization presents:

- noisy input (headers, tags,...)
- no guarantee of linguistic well-formedness
- mixed properties of oral and written language
- multi-topic messages

High-quality, general-purpose NLP tools cannot deal properly with such a bulk of asystematic differences from standard texts. This implies a barrier for approaches to text summarization that have proven successful in more standard genres, because they are crucially relying on the output of such tools.

As a consequence, very little work has been done on quality e-mail summarization. (Tzoukermann et al., 2001) aim to capture the gist of e-mail messages by extracting salient noun phrases, using a combination of machine learning and shallow linguistic analysis. (Nenkova and Bagga, 2003) apply standard multi-document summarization techniques to produce written extracts of e-mail threads.

Considering e-mail summarization problems and the environment within PETRA project, summaries produced by CARPANTA have the following properties:

oral output by telephone,

indicative summaries just give a hint of the content, to meet the severe restrictions of length imposed by the oral format,

coherent because the summary cannot be revised as easily as written ones, (thus excluding list-of-words approach),

extractive due to limitations for general-purpose NLP tools,

knowledge-intensive combining analysis at different linguistic levels, IR techniques and IE strategies specific for e-mail, in order to build a robust system that is also capable of producing deep analyses.

3 Architecture of the System

As can be seen in Figure 1, CARPANTA is highly modular, which guarantees portability to other languages. The core processing stream is formed by language-independent strategies, while e-mail specific knowledge is in autonomous modules that can be updated and switched to address concrete necessities (different languages, restricted domains).

In addition to general-purpose NLP tools, the following e-mail specific resources were developed:

- a classification where each kind of e-mail is associated to its most adequate summary and summarization strategy (language-independent) (seen in Table 1)
- bags of words and expressions that signal different kinds of e-mail specific contents (language-dependent)
- strategies to deal with these anchors and their associated content (language-independent)

The process for e-mails to be summarized is described in what follows.

Parse e-mail format. Messages undergo a pre-processing to identify headers, greetings, visit cards, quoted text, and the body of text, which is further analyzed.

Linguistic analysis. First, the body of text is analyzed morphosyntactically (Atserias et al., 1998a) and chunks are identified (Atserias et al., 1998b). Then, discourse chunks, signalled by punctuation and discourse markers, are found (what we call *segments*). Finally, the salience of non-empty words is calculated according to the frequency of occurrence of their lemma.

Textual analysis. Three different kinds of textual relevance have been distinguished: lexic, structural and subjective. For each of these three aspects of e-mails, a global reliability score is obtained, taking into account how well each kind of information distinguishes relevant and non-relevant pieces of the e-mail. Then, relevance is also calculated with respect to meaning units, basically, discourse segments. Lexic relevance of a segment is directly proportional to the amount of frequent words in the segment and inversely proportional to the length of the segment. Structural relevance is assigned as a result of the interpretation of discursive relations between segments and between a segment and the whole text, by means of the information associated to discourse markers. Finally, subjective relevance is found when the segment contains any of a list of lexical expressions signalling subjectivity.

```
if strong genre evidence
   if strong linguistic evidence
       textual + documental
   else
       if evidence for a single genre
           specific strategy
           (list, question, attachment)
       else combination of genres
else
    if strong textual evidence
        textual
    else lead
```

Figure 2: General schema followed by classification rules.

Documental analysis. Key words and expressions signalling information specific of e-mail (e.g., appointment, list, etc.) are detected by simple IE techniques, basically, pattern-matching.

As a result of linguistic, textual and documental analysis, a set of meaning units is produced at different linguistic levels: words, chunks, segments and sentences, but also lines and paragraphs. Each unit is assigned a complex relevance score, one for each kind of information that is taken into account. Values for lexical, structural and subjective relevance are continious, ranging from 0 to 1. Each unit is also assinged a binary relevance score for each kind of e-mail specific information, 1 if there is any clue signalling that kind of information in the unit, 0 if there is none.

Classification The most adequate summarization strategy is determined by taking into account the characterizing features of each e-mail, as provided by the analysis module. The general schema followed by classification rules can be seen in Figure 2, Table 1 shows the relation between e-mail features and summarization strategies.

Summarization Then, the chosen summary is produced. Different kinds of summaries are described in Table 1.

4 Results and Discussion

To tune and evaluate the performance of the system, the automatic summaries produced

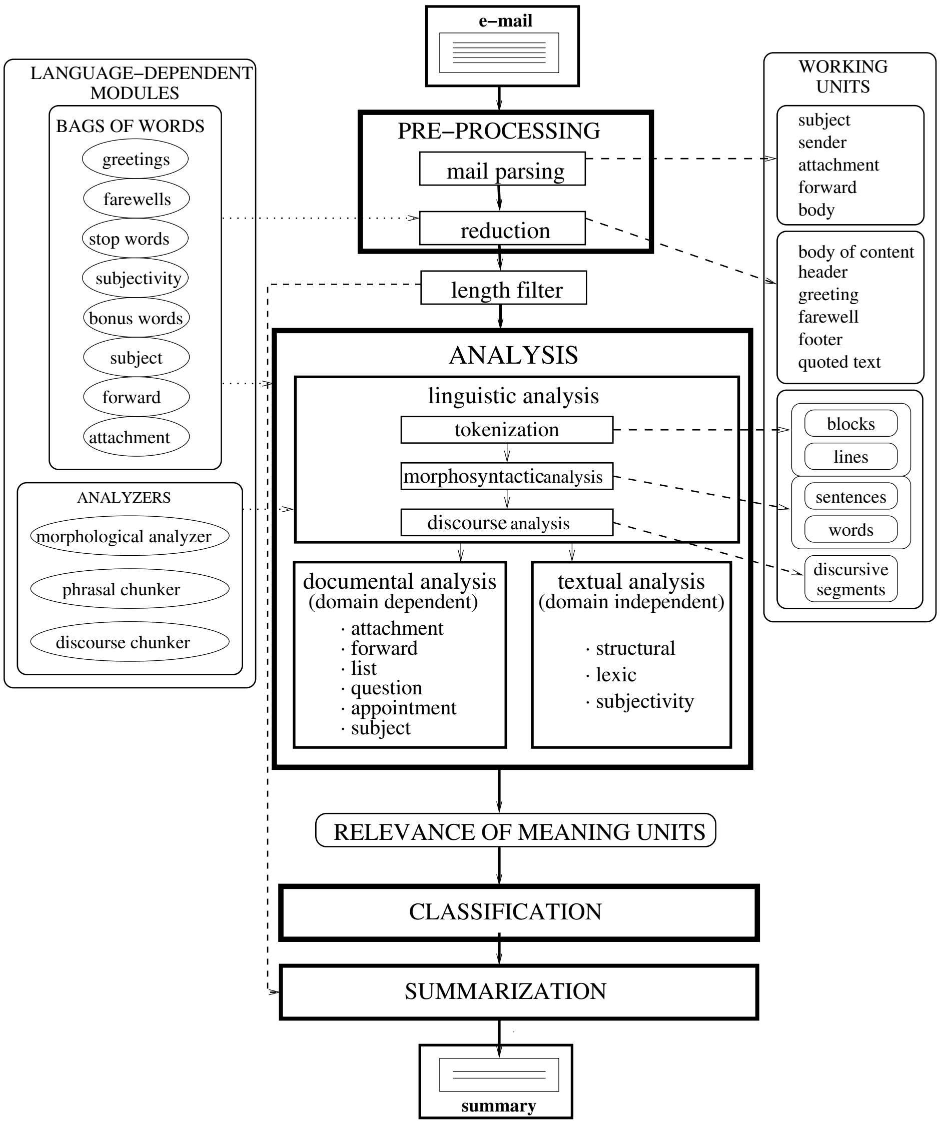

Figure 1: Architecture of CARPANTA.

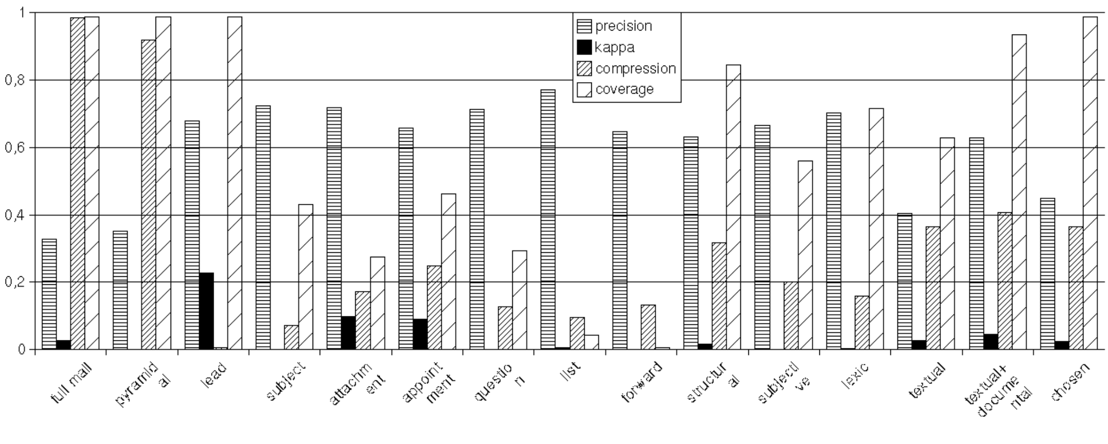

Figure 3: Main features of the performance of different summarization strategies: average length with respect to the original e-mail text (compression), coverage over the test collection, and kappa agreement and precision by comparison to human summaries.

summarization approach	summary	textual features	documental features
full mail	whole e-mail text	short (<30 words)	
pyramidal	first compressed paragraph	none is relevant	none is relevant
lead	first compressed sentence	none is relevant	none is relevant
subject	subject of e-mail	strong lexical relevance	subject is relevant
appointment	segment stating appointment	none is relevant	evidence of appointment
attachment	segment describing attachment	none is relevant	evidence of attachment
forward	segment describing forward	none is relevant	evidence of forward
question	segment with question	none is relevant	question mark
list	segment preceeding the list	none is relevant	list
lexic	segment with most relevant lexic	strong lexical relevance	none is relevant
structural	most structurally salient segment	strong structural relevance	none is relevant
subjective	segment with subjectivity evidence	strong subjective relevance	none is relevant
textual	most relevant segment summing all textual evidence	none is salient	none is salient
textual + documental	most relevant segment summing all textual and documental evidenc	none is salient	none is salient

Table 1: Classification of summaries, characterizing features and summarization strategies.

were compared with summaries produced for 200 e-mails by 20 potential users of the system, with a minimum of 2 different human summaries for each e-mail. Agreement between judges ranged from $\kappa = -.37$ to $\kappa = 1$, with a mean of $\kappa = .47$, which indicates that agreement is far beyond chance, but also that the task of e-mail summarization is somewhat fuzzy for users.

The goodness of automatic summaries was calculated by comparison with the corresponding human summaries, results can be seen in Figure 3. For each e-mail, automatic summaries were obtained using all of the summarization strategies applicable, based on linguistic information (*lexical*, *structural*, etc.), on e-mail specific information (*appointment*, *attachment*, etc.) in both (*textual and documental*) or applying baseline strategies, like having the first line or paragraph as the summary.

Human and automatic summaries were compared by κ agreement and by precision at discourse unit level. Agreement between human and automatic summaries was very low in terms of κ (average $\kappa = .02$), but evaluation metrics more usual for summarization, like precision with respect to human summaries, reached 60% average, which is the state of the art for automatic text summarization.

Results show that simple methods, like taking the first line of the e-mail (*lead*) offer very good results, but, in general, summaries exploiting e-mail specific knowledge (*list, appointment*) can improve on this baseline. However, these kinds of e-mail present very low coverage. The strategy combining general linguistic and e-mail specific knowledge (*textual and documental*) yields a good balance between coverage and precision.

Finally, results concerning the chosen summary show that there is still room for improvement within the classification module, since most of the alternative summaries present higher precision rates than the chosen one.

5 Conclusions and Future Work

We have presented CARPANTA, an e-mail summarization system that applies a knowledge-intensive approach to obtain highly coherent summaries, targeted to guarantee understandability in delivery by phone. Results indicate that the classification module has to be improved. Given the highly modular architecture of CARPANTA, adaptation to other languages has a very low cost of development, provided the required NLP tools are available.

References

J. Atserias, J. Carmona, S. Cervell, L. Màrquez, M. A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. 1998a. An environment for morphosyntactic processing of unrestricted spanish text. In First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.

J. Atserias, I. Castellón, and M. Civit. 1998b. Syntactic parsing of unrestricted spanish text. In First International Conference on Language Resources and Evaluation, Granada. LREC.

A. Nenkova and A. Bagga. 2003. Facilitating email thread access by extractive summary generation. In RANLP 2003.

E. Tzoukermann, S. Muresan, and Judith L. Klavans. 2001. Gist-it: Summarizing email using linguistic knowledge and machine learning. In ACL-EACL'01 HLT/KM Workshop.