Carpanta eats words you don't need from e-mail

Laura Alonso[*], Bernardino Casas[†], Irene Castellón[*]
Salvador Climent[‡], Lluís Padró[†]

[*]GRIAL
Dept. de Lingüística General
Universitat de Barcelona
{lalonso,castel}@fil.ub.es

[†]TALP Research Center
Software Department
Universitat Politècnica de Catalunya
{bcasas,padro}@lsi.upc.es

[‡]Estudis d'Humanitats
i Filologia
Universitat Oberta de Catalunya
scliment@uoc.edu

Resumen: Presentamos Carpanta, un sistema de resumen automático de correo electrónico que aplica técnicas de conocimiento intensivo para obtener resúmenes coherentes. El uso de herramientas de PLN de amplia cobertura garantiza la robusteza y portabilidad del sistema, pero también se explota conocimiento dependiente de lengua y dominio. Carpanta ha sido evaluado por comparación con un corpus de resúmenes confeccionados por jueces humanos, con resultados satisfactorios.
Palabras clave: Resumen Automático, Correo-e

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, 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.
Keywords: Automatic Text Summarization, E-mail

1 Introduction

We present Carpanta, the e-mail summarization system within project Petra, funded by the Spanish Government (CICyT TIC-2000-0335). The global goal of the project is to develop an advanced and flexible system for unified message management, which enhances the mobility, usability and confidentiality levels of current systems, while keeping compatibility with main nowadays computer–phone integration platforms. 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".

The project includes three work lines:

1. **Integration** of phone, internet and fax services.

2. Development of advanced **oral interfaces** based on speech recognintion and understanding, speech synthesis, and speaker verification.

3. Intelligent **information management** through the use of Natural Language Procesing (NLP) techniques for text classification and summarization, as well as for information retrieval. This task includes the subgoals of advanced Named Entity recognition and correference resolution, document filtering, categorization and retrieval, and text summarization, being this last issue 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 its modular architecture, with a language-independent core and separated modules exploiting language-dependent knowledge.

The rest of the paper is structured as follows: first, NLP problems specific to e-mail summarization are described. Section 3 presents our approach to e-mail analysis and summarization, then, the architecture of the system is sketched. Section 5 introduces the evaluation by comparison with a human-made golden standard, results can be seen in Section 6. We finish with some conclusions and future work.

2 Problems of e-mail summarization

Automatic Summarization has become in last years an active line of research. Initially reduced to a textual, monolingual, single-document condensation task, it has evolved for covering a wide spectrum of tasks and applications, each presenting common points with the general task of summarization, but also idiosyncratic problems. For e-mail summarization, the major problems are:

- noisy input (headers, tags,...)
- linguistic well-formedness is far from guaranteed
- properties of oral and written language
- multi-topic messages

Many scholars have studied relevant aspects of the e-mail register. They have mainly focused on the similarities and differences between oral language and texts (Yates and Orlikowski, 1993; Ferrara, Brunner, and Whittemore, 1990) as well as in brand new intentionally-expressive devices, such as previous-message cohesion (Herring, 1999), visual devices (Fais and K., 2001), simplified registers (Murray, 2000) or internet-users vocabulary (Alonso, Folguerà, and Tebé, 2000). Nevertheless, they disregard a factor that is important in the e-mail register: as the user often writes not much reflectively, texts contain many non-intentional language mistakes.

In a recent study, Climent et al. (2003) argue that, for their universe of study, more than 10% of the text in emails are made of either non-intentional errors, intentional deviations of the written standards, or specific terminology. For Spanish, 3.1% of the words contain either performance or competence errors, another 3.3% are either language-shifts or new forms of textual expressivity (such as ortographical innovations or, specially, systematic non-accentuation), and another 4.2% consist of specific terminology -thus words usually missing from many system's lexicons.

In any case, such a bulk of asystematic differences from standard texts implies a barrier for high-quality, general-purpose NLP tools. As a consequence, very little work has been done on quality e-mail summarization. Tzoukermann, Muresan, and Klavans (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.

3 Approach

As presented in the general environment of PETRA, the output of the summarization system is a telephone message. Given the severe restrictions in summary length imposed by the oral format, we chose to provide *indicative* summaries that give a hint of the content, instead of longer, informative summaries, which tend to synthesize most of the relevant information.

Moreover, the *understandability* of the message has to be much higher than it is necessary for written summaries, because the summary cannot be revised as easily in case the user cannot understand properly. This excludes a list-of-words approach, because a list of noun phrases is too incoherent to be easily understood by phone.

Finally, we have taken a *knowledge-intensive* approach to summarization, combining analysis at different linguistic levels, IR techniques and information extraction strategies specific for e-mail. As a consequence, robustness is guaranteed by domain-independent analysis, while the systematicities that can be found in e-mail are exploited in a specific, deeper level of analysis.

It must be said that, due to limitations in NLP capabilities, summaries were not generated, but built by *extraction* of fragments of the original e-mail, which supposes a shortcoming with respect to coherence. Nevertheless, in contrast to usual extractive summarization, the size of the extracted fragments was not based on ortography, that is to say, we did not extract sentences, but discourse-motivated *segments*.

Discursive segments are complete linguistic structures, no smaller than a phrase and no bigger than a sentence, bearing the necessary propositional content to constitute a felicitous sentence, even if a certain kind of supplementation from a matrix structure is needed, exploiting the same kind of mechanisms that apply for in the intrepretation of *fragments*. Moreover, as discussed in Alonso and Castellón (2001), the constitution of a segment must not cause ungrammaticality or infelicity in the surrounding discourse. Discourse segments are identified by an automated discourse chunker (see next Section). Well-formedness of the extracted fragments of text is guaranteed by extracting both the selected segments and their eventual matrix structures, usually, the core of the sentence.

4 Architecture of the System

As can be seen in Figure 1, CARPANTA is highly modular, which guarantees portability to other languages. Moreover, this modular architecture spares much implementation efforts, since it takes advantage of pre-existing NLP tools, like the basic analyzers.

E-mail specific knowledge has different status within the system, so that language-dependent modules can be updated and switched to address concrete necessities (different languages, restricted domains), while language-independent strategies form part of the core processing stream. 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)
- bags of words and expressions that signal different kinds of e-mail specific contents (language-dependent):
  - greetings, farewells,
  - reply, forward, attachment
  - bags of words signalling different kinds of relevance: personal involvement of the writer in the message, information exchange; also lack of relevance.
- strategies to deal with anchors and associated content (language-independent)

To parse e-mail format, messages undergo a pre-processing that identifies pieces like headers, greetings, visit cards and, of course, the body of text. E-mails that are an answer to previous ones undergo a special pre-processing to determine whether the text of the previous message should be taken into account as constituting the summary.

4.1 Analysis

The analysis of the e-mail combines domain-independent and domain-dependent knowledge. A basic analysis gathers information about the documental, textual and linguistic structure of the message, whereupon e-mail specific analysis machinery is applied.

In the first place, basic document units, lines and paragraphs, are found. These units can be used when the linguistic structure of the text is not informative enough or when there is no other segmentation method available, for example, when there is no chunker for the language. This step is specially error-prone, because the meaning of the symbol for a newline is highly ambiguous, as it is totally subject to personal style.

As the basis of the textual analysis, a morphosyntactic process is applied. In this step, punctuation marks and lexical tokens are recognized and POS tags are assigned to words (Carmona et al., 1998). Also, a partial syntactical analysis is carried out (Atserias, Castellón, and Civit, 1998), which recognizes noun, prepositional and adjectival phrases and complex verbal forms. Then, discourse chunks, signalled by punctuation and discourse markers, are found by a discourse segmentation grammar. This discourse segmentation grammar also establishes the relative relevance and shallow coherence relations between discourse segments by resorting to a discourse marker lexicon (Alonso, Castellón, and Padró, 2002). Finally, the salience of non-empty words is calculated according to the frequency of occurrence of their lemma. It has to be noted that the lack of well-formedness of e-mails increases the error rate of these general-purpose analysis tools far beyond their usual performance level.

The documental analysis concerns the identification of e-mail specific clues and their accompanying information, by simple IE techniques like pattern-matching.

The output of this module is the set of meaning units at different linguistic levels: words, chunks, segments and sentences. These co-exist with meaning units at document level, lines and paragraphs. Each unit is assigned a relevance score according to the amount and kind of relevance encountered in it. Values for textual relevance are continuous from 0 to 1, values for documental, e-mail specific knowledge are binary, recording the presence of any clue in a segment. Each kind of textual relevance is assigned a score for global reliability, based on the strength of the evidence found.

Three different kinds of textual relevance have been distinguished: lexic, structural and subjective. 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 be-

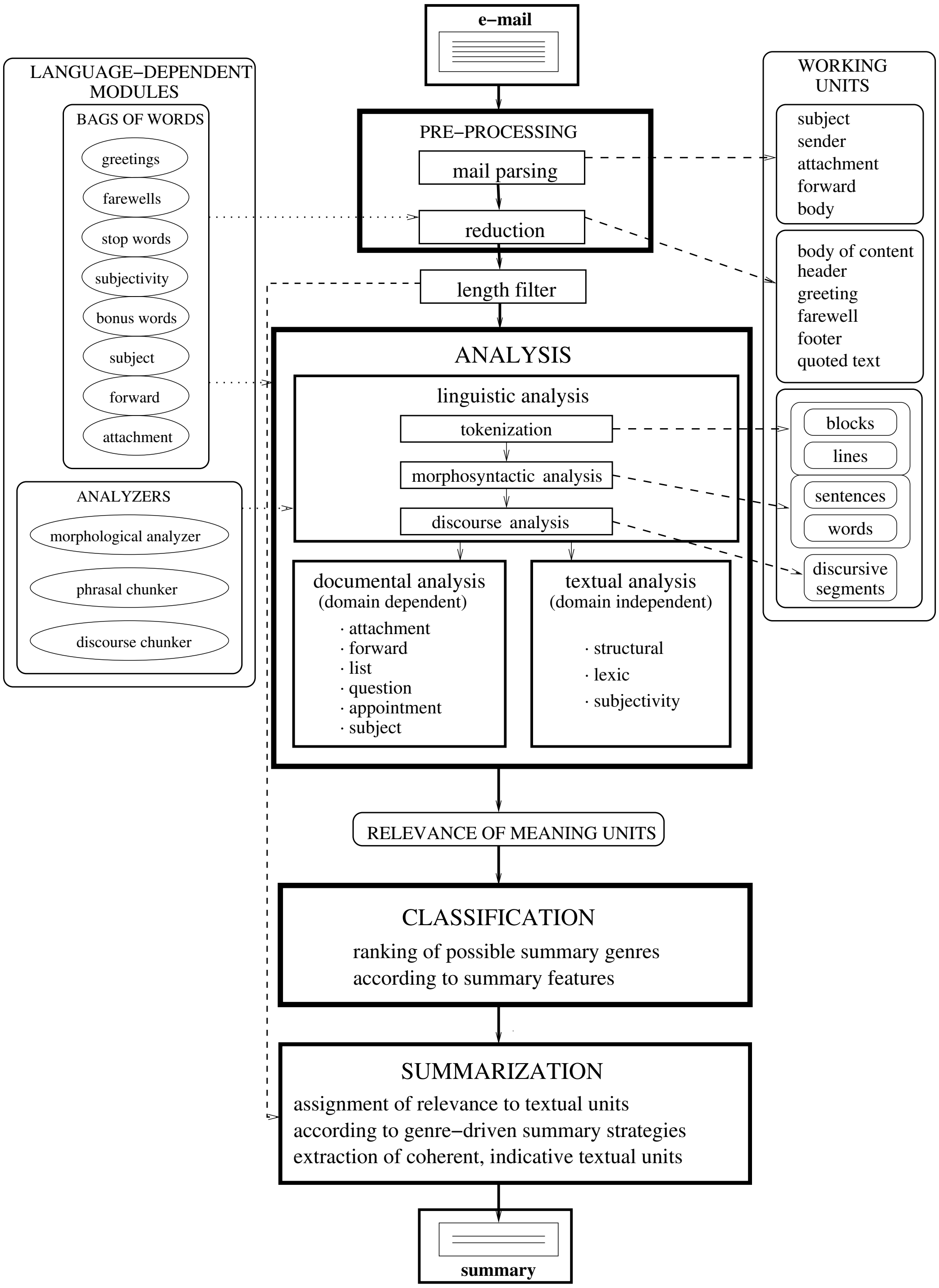

Figure 1: Archictecture of CARPANTA.

tween segments and between a segment and the whole text, by means of the information associated to a set of discourse markers. Finally, subjective relevance is found when the segment contains any of a list of lexical expressions signalling subjectivity.

4.2 Classification and Summarization

The classification module determines the most adequate summarization strategy by taking into account the characterizing features of each e-mail, provided by the analysis module. It works by applying hand-made rules that follow the general scheme described in Figure 2.

Then, the chosen summary is produced by the summarization module. 13 different summarization strategies are provided, a rough relation between e-mail features and summarization strategies can be seen in Table 1.

5 Evaluation

To tune and evaluate the performance of the system, the automatic summaries produced were compared with summaries produced by potential users of the system. 200 e-mails were summarized by 20 judges, so that each e-mail was summarized by at least 2 judges. The average e-mail length was 340.7 words, 14.6 sentences and 9.8 paragraphs[1]. Of the 200 e-mails, 36% contained more than one pre-defined documental structure, like lists, questions, etc.; 41% presented none.

Judges were instructed to mark those words in the e-mail text which they would find useful as a summary, provided by phone, to get a general idea of the content of the message. No guidelines were provided as to the length or type of the textual fragments to be marked. Since the intended goal of e-mail summarization is ill-defined, judges produced both a representation of the goal and the gold standard to evaluate it.

As a consequence of this double purpose of human judgements, only 20% of the judged e-mail was used for evaluation (test corpus), the rest was used for characterizing the features of the intended summaries and tuning the system (development corpus). For example, targeted summary length was determined as the average length of human summaries (26 words). However small the test corpus may seem (40 e-mails), it supposes a significant enhancement upon previous evaluation of automatic e-mail summaries, like Tzoukermann, Muresan, and Klavans (2001), who only used 8 e-mails.

The *kappa* measure (Landis and Koch, 1977) was used to calculate pairwise agreement between judges. Kappa is a better measurement of agreement than raw percentage agreement because it factors out the level of agreement which would be reached by random. When there is no agreement other than what would be expected by chance, $k = 0$, when agreement is perfect, $k = 1$. Additionally, word-based measures, like unigram and bigram overlap, were used to account for equivalences in informativeness between human and automatic summaries (Mani, 2001).

The obtained kappa values for agreement between judges ranged from 0.36 to 1, with a mean of 0.75 and a standard deviation of 0.17. Following (Carletta, 1996), we can consider that kappa values above 0.7 indicate good stability and reproducibility of the results, so it can be said that it is possible to discriminate a good e-mail summary from a bad one, and that it is even possible to determine the best summary for a given e-mail.

The goodness of automatic summaries was calculated as the agreement with the corresponding human summaries, at word level. Instead of the usual *recall* and *precision* measures for comparing an automatic summary with a gold standard, the kappa measure was used, precisely because, as said before, it factors out chance agreement.

As a global measure of the system's performance, we calculated the effect of considering the system as a human judge more, with respect to average kappa agreement. Taking the 20% of the corpus left apart for summarization, we obtained that the average kappa agreement between human judges was 0.74, and it decreased to 0.54 when the system was introduced as a judge more. This indicates that the system does not as well as human judges, but still, a kappa value bigger than 0.4 indicates moderate agreement.

Concerning informativeness, unigram overlap between summaries from different judges reached an average of 0.44, and bigram overlap amounted to 0.36. When automatic summaries were compared to the human gold standard, overlap never reached 0.4, and in some cases it didn't even amount to 0.2. Therefore, these measures did not allow to detect significant differences in summarization strategies.

6 Results and Discussion

Figure 3 shows the results of comparing automatic summaries against human-made summaries of the 40 e-mails reserved for evaluation. For each e-mail, automatic summaries were obtained using all of the summarization strategies applicable, for example: *lexic*, *structural*, *appointment*, *attachment*, etc.

[1]The number of sentences and paragraphs is approximate, due to the high asistematicity of the usual cues for segmentation at these levels (full stops, carriage returns) in e-mail texts.

if e-mail body is short
  then *full mail*
else if strong e-mail specific evidence
  if no strong textual evidence
    if one single e-mail specific evidence
      then single *genre-driven* (*subject, appointment, attachment, etc*)
    else ponderated multiple *genre-driven* (*textual + documental*)
  else *textual + documental*
else if strong textual evidence
  if one single textual evidence
    then single *textual*
  else *textual*
else *pyramidal*

Figure 2: Outline of the rules for classification of e-mails, to determine the best-suited summarization strategy taking into account the e-mail features.

summarization approach	**summary**	**textual features**	**documental features**
full mail	whole e-mail text	short (<30 words)	
pyramidal	first paragraph in e-mail with no irrelevant segments	none is relevant	none is relevant
subject	subject	strong lexical relevance	subject is relevant
appointment	segment with time of event of appointment	none is relevant	lexical evidence of appointment
attachment	segment with description of statement of attachment	none is relevant	lexical evidence of attachment
forward	segment with description of statement of forward	none is relevant	lexical evidence of forward
question	segment with question	none is relevant	question mark
list	segment preceeding the list, first segment of items	none is relevant	list
lexic	segment containing most relevant lexic	strong lexical relevance	none is relevant
structural	segment most salient structurally	strong discourse structural relevance	none is relevant
subjective	segment most salient subjectivity	strong subjective relevance	none is relevant
textual	most relevant segment summing all textual relevance evidence	none is salient	none is salient
textual + documental	most relevant segment summing textual and documental relevance	none is salient	none is salient

Table 1: Pre-established kinds of summaries, characterizing features of each kind and associated summarization strategies.

Results show average statistics of the comparisons between human and automatic e-mails grouped by the kind of strategy applied, which permits a separate evaluation of different kinds of summaries and also an evaluation of the best summary choice. Due to the small size of this evaluation corpus, some of the summarization strategies did not apply, and are not represented in the evaluation, like *list*, *attachment*, *forward* or *subject*. However, they were found in the training corpus, and performance for these strategies is very much comparable to that of other e-mail specific strategies, like *appointment* or *question*.

The main quality measure of evaluation was kappa agreement, calculated between automatic summaries and every human summary available for that e-mail. Unigram and bigram overlap are also provided. Moreover, compression rate and coverage are also calculated for every strategy. Compression rate expresses the length of the summary as a ra-

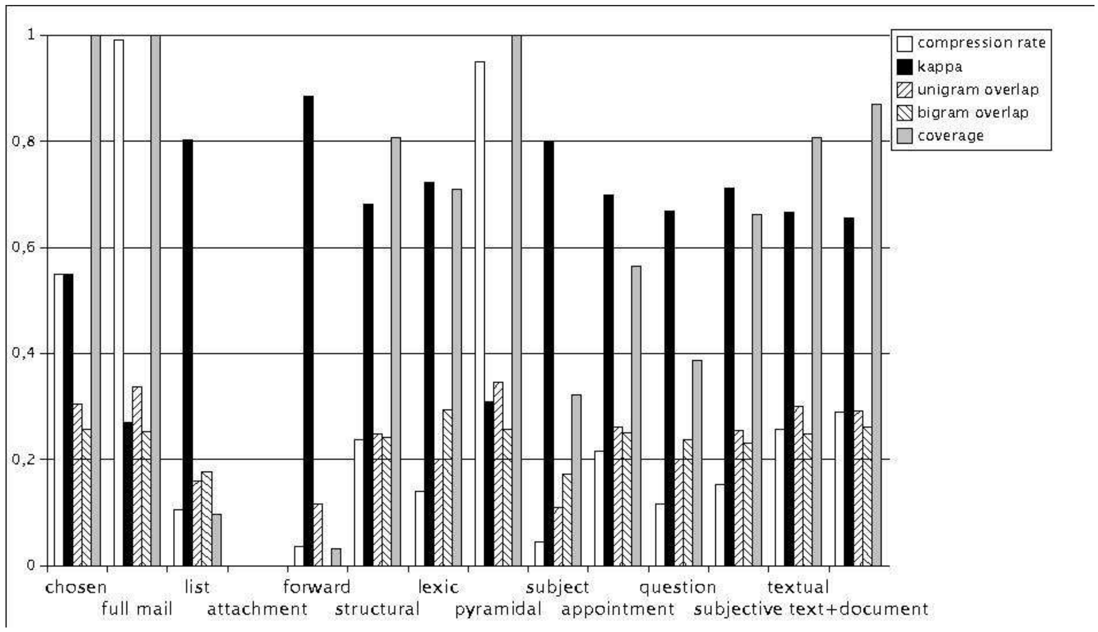

Figure 3: Main features of the performance of different summarization strategies: the average quality of each strategy is described by kappa agreement and the word-based unigram and bigram overlap. Compression rate expresses the ratio of the length of the original message that the summary represents, and coverage accounts for the percentage of mails in the corpus that can be summarized by a given strategy.

tio of the length of the original message, so that longer summaries have a higher compression rate. Coverage accounts for the percentage of mails in the corpus that can be summarized by a given strategy, for example, *full mail*, *pyramidal* and of course *chosen* are the only strategies to provide a summary for all the e-mails in the corpus.

It is shown that a knowledge intensive approach yields better summaries than simpler methods, like taking the first paragraph of the e-mail. It can be seen that *pyramidal* strategy yields a very bad balance between summary length and agreement with judges, almost equalling *full mail* approach. Therefore, and opposed to usual kinds of summarization, location in the e-mail cannot considered as feature for relevance.

In general, summaries exploiting e-mail specific knowledge show higher kappa agreement with human judgement than textual-based ones, but the latter present a much higher coverage (0.77 mean coverage for textual-based summaries in contrast to 0.23 for e-mail specific ones). Indeed, textual-based strategies apply for most of the collection of e-mail, while not every message contains e-mail specific clues that have been previously systematized. The strategies *textual* and *textual + documental* suppose a compromise between precision and coverage. As can be expected, they present a very good relation between summary length and agreement with human summaries.

It must be said that very simple techniques, like taking the segments with the most frequent words in text (strategy *lexic*) or those asking a question (strategy *question*) also yield very good results. This indicates that a better account of how each kind of evidence contributes to obtain a good summary will improve the strategies combining different kinds of information, as is the case for *textual* and *textual + documental*.

Finally, results concerning the chosen summary show that there is still room for improvement within the classification module. The final summary, chosen from all summaries produced for a certain e-mail, presents good agreement with the summaries made by humans, but the average length is quite high.

7 *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. The performance of the system has been evaluated with a corpus of human-made e-mail summaries, reaching a level of agreement with users close to agreement between human judges. However, results indicate that the classification module has to be improved, which will be done by manually incrementing the rules and by applying machine learning techniques.

Given the highly modular architecture of CARPANTA, adaptation to other languages has a very low cost of develpment, provided the required NLP tools are available. Indeed, enhancements for Catalan and English are under development.

Future work in our system should include modules that enable for automatic normalization and correction of input texts. (Climent et al., 2003) suggest that there's special need for modules of: (a) punctuation recovery, (b) accent recovery, (c) spelling-mistake correction, and (d) terminological tuning according to users' profiles.

Acknowledgements

This research has been conducted thanks to a grant associated to the X-TRACT project, PB98-1226 of the Spanish Research Department. It has also been partially funded by projects HERMES (TIC2000-0335-C03-02), PETRA (TIC2000-1735-C02-02), and by CLiC (Centre de Llengutatge i Computació).

References

Alonso, A., R. Folguerà, and C. Tebé. 2000. Del tecnolecte al sociolecte: consideracions sobre l'argot tècnic en català. *I Jornada sobre Comunicació Mediatitzada per Ordinador en Català (CMO-Cat)*. Universitat de Barcelona.

Alonso, Laura and Irene Castellón. 2001. Towards a delimitation of discursive segment for natural language processing applications. In *First International Workshop on Semantics, Pragmatics and Rhetoric*, Donostia - San Sebastiàn, November.

Alonso, Laura, Irene Castellón, and Lluís Padró. 2002. Design and implementation of a spanish discourse marker lexicon. In *SEPLN*, Valladolid.

Atserias, Jordi, Irene Castellón, and Montse Civit. 1998. Syntactic parsing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation*, Granada. LREC.

Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carmona, Josep, Sergi Cervell, Lluís Màrquez, M. Antònia Mart, Lluís Padró, Roberto Placer, Horacio Rodríguez, Mariona Taulé, and Jordi Turmo. 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Climent, S., P. Gispert-Saüch, J. Moré, A. Oliver, M. Salvatierra, I. Sànchez, M. Taulé, and Ll. Vallmanya. 2003. Machine translation of newsgroups at the uoc. evaluation and settings for language control. *Journal of Computer-Mediated Communication*. in press.

Fais, L. and Ogura K. 2001. Discourse issues in the translation of japanese e-mail. *Proceedings of the Pacific Association for Computational Linguistics, PACLING 2001.*

Ferrara, K., H. Brunner, and G. Whittemore. 1990. Interactive written discourse as an emergent register. *Written Communication*, 8:8–34.

Herring, S. 1999. Interactional coherence in cmc. *Journal of Computer-Mediated Comunication*, 4(4). special issue on Persistent Conversation.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, March.

Mani, Inderjeet. 2001. *Automatic Summarization*. Nautral Language Processing. John Benjamins Publishing Company.

Murray, D. E. 2000. Protean communication: the language of computer-mediated communication. *Tesol Quarterly*, 34(3):397–421.

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

Yates, J.A. and W.J. Orlikowski. 1993. Knee-jerk anti-loopism and other e-mail phenomena: Oral, written, and electronic patterns in computer-mediated communication. Working Paper 3578-93, MIT Sloan School. Center for Coordination Science Technical Report 150.