Combining heterogeneous knowledge sources in e-mail summarization

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

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, and the contribution of each kind of information to summary goodness is assessed.

1 Introduction

We present CARPANTA, the e-mail summarization system within project PETRA, funded by the Spanish Government (CICyT TIC-2000-0335). 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.
2. Development of advanced **oral interfaces**.
3. Intelligent **information management** through the use of Natural Language Procesing (NLP) techniques for information retrieval, text classification and summarization, being this last issue specially relevant for oral interfaces to electronic mail systems.

CARPANTA is the summarization module within PETRA. Its function is to summarize incoming e-mail, so that it can be readily delivered to the user by phone. It is currently working for Spanish, but portability to other languages is guaranteed by its modular architecture, which allows re-usability of already existing tools. Its core processing stream is language-independent, and language-dependent knowledge is provided by separated modules that can be easily integrated in the system.

The rest of the paper is structured as follows: in Section 2, the main aspects of e-mail summarization for telephone delivery are presented, and the architecture of the system is sketched. Section 3 describes in detail the basic components of the system. Section 4 presents an evaluation of the performance of CARPANTA by comparison with a human-made golden standard, highlighting how each kind of information contributes to obtaining good summaries. We finish with some conclusions and future work.

2 Aspects of e-mail for telephone summarization

A commonly assumed classification of the aspects that influence text summarization (Sparck-Jones 99) distinguishes *input*, *purpose* and *output* aspects. Input and output aspects have played a crucial role in the design of CARPANTA.

Input. e-mail register presents many idiosincracies that escape the rules of the standard language usage (Yates & Orlikowski 93; Ferrara *et al.* 90; Herring 99; Fais & Ogura 01; Murray 00; Alonso *et al.* 00). In a recent study (Climent *et al.* 03), it is argued that more than 10% of the text in e-mails is made of either non-intentional errors, intentional deviations of the written standards, or specific terminology. Therefore, email-oriented NLP has to be robust enough to "*gracefully degrade - rather than crash - when confronted with unexpected data*" (Stede 03, pg. 1). More concretely, CARPANTA has to deal with:

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

Output. the format of CARPANTA's summaries is a telephone message. The oral format imposes severe restrictions in summary length. Therefore, CARPANTA creates summaries that are *indicative* of the e-mail content, in contrast with informative summaries, which tend to synthesize most of the relevant information. In addition, since the summary cannot be revised as easily as in written format, a highly *coherent* text must be provided. Previous work on e-mail summarization has mainly focussed in informativeness; for example, (Tzoukermann *et al.* 01) aim to capture the gist of e-mail messages by extracting salient noun phrases, using a combination of machine learning and shallow linguistic analysis. In contrast, CARPANTA is not only concerned with content, but also with the form of the summaries.

As follows, coherence of the summaries is a compelling feature in CARPANTA, although the main objective is robustness, that is to say, that a summary is provided for every incoming e-mail. In order to satisfy both these requirements, CARPANTA applies a knowledge-intensive approach to summarization based in a combination of robust analysis tools, integrating linguistic analyzers at different levels, IR techniques and information extraction strategies specific for e-mail.

As can be seen in Figure 1, the architecture of CARPANTA guarantees robustness with a domain-independent processing stream based on shallow linguistic analysis, described in Section 3.1.

As developed in Section 3.2, the systematicities of the domain are also exploited, but this deeper knowledge is not robust in terms of coverage or reliability. This is why CARPANTA does not crucially rely on domain-specific knowledge to produce a summary, although it integrates it when available, as is reflected in its architecture.

As a result of the basic linguistic analysis, each e-mail is broken down into meaning units. Each of these units is assigned a relevance score according to the amount and kind of relevance encountered in it. Values for basic linguistic (*textual*) relevance are continuous from 0 to 1. Additionally, each kind of textual relevance is assigned a score for global reliability, based on the strength of the evidence found for that kind of relevance. Values for e-mail specific (*documental*) knowledge are binary, recording the presence of any e-mail specific clue in each meaning unit.

Once an e-mail has been analyzed, it is classified by its characterizing features, in order to determine the optimal summarization strategy to be applied, as exposed in Section 3.3. Summarization strategies, seen in Table 1, range from very specific to very general, so that highly targeted summaries can be provided when enough information is available, but some kind of summary is always produced, even when there is no useful information on the e-mail.

The resulting summaries are formed by one or more literal fragments of the original e-mail text, the most common method to build summaries in automated text summarization systems, because the state of the art in NL Generation or Regeneration yields even more incoherent texts. Nevertheless, in contrast to usual extractive summarization, the extracted fragments are discourse-motivated, instead of based on ortography.

3 Main Components of CARPANTA

3.1 Textual Analysis

The output of the textual analysis is a set of meaning units at different linguistic levels: words, chunks, discursive segments and sentences. These co-exist with meaning units at document level, lines and paragraphs. Whenever it is possible, discursive segments are taken as the basic meaning unit to which relevance is assigned. However, when this is not possible, lines or paragraphs are taken as meaning units.

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.* 98). Also, a partial syntactical analysis is carried out (Atserias *et al.* 98), which recognizes noun, prepositional and adjectival phrases and complex verbal forms.

Then, discourse segments, signalled by punctuation and discourse markers, are found by a discourse segmentation grammar. Discourse 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* (Ginzburg & Sag 00). Moreover, the constitution of a segment must not cause ungrammaticality or infelicity in the surrounding discourse (Alonso &

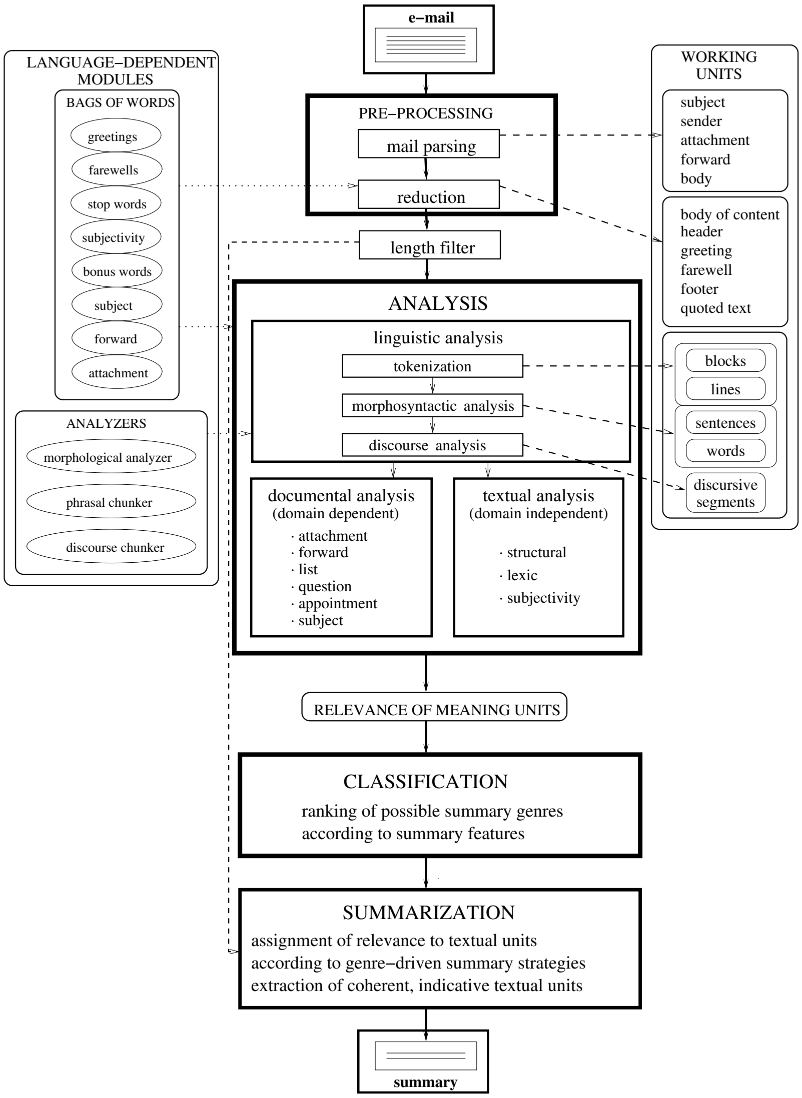

Figure 1: Archictecture of CARPANTA.

Castellón 01). The relative relevance and shallow coherence relations between discourse segments is established by resorting to a discourse marker lexicon (Alonso *et al.* 02).

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[1] 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. Finally, subjective relevance is found when the segment contains any of a list of 120 expressions signalling subjectivity.

[1]Frequency of words is calculated after stopwords have been removed and lemmatization has been performed.

3.2 Documental Analysis

The documental analysis concerns the identification of e-mail specific clues and their accompanying information, by simple IE techniques like pattern-matching. These clues are lists of regular expressions or words, either lemma or form, that signal different kinds of e-mail specific content.

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 summary text.

Most of the clues to carry out the documental analysis and the parsing of e-mail format are language-dependent; the following lists were created specifically for Spanish (the number of items for each list is provided):

- greetings (21), farewells (26), might-be farewell formulas (3), meeting formulas (27)
- forward (2), attachment (15)
- bonus words (87) and stigma words (5)
- list (7) and quote marks (3), topic shifts (9)

3.3 Classification and Summarization

Taking into account the characterizing features of each e-mail, which are provided by the analysis module, the classification module determines the most adequate summarization strategy within a choice of 13. The scheme followed by the classification rules is described in Figure 2.

if strong e-mail specific evidence
    if strong textual evidence
        then *textual + documental*
    else if one single e-mail specific evidence
        then single *genre-driven*
        (*subject, appointment, attachment, etc*)
    else ponderated multiple *genre-driven*
    (*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.

The general aim of the classification module is to determine the most adequate summarization strategy that can be applied to each e-mail with a reliable level of confidence given its characterizing features. The specificity of the chosen summarization strategy is proportional to the specificity of the characterizing features. When no informative features are provided for an e-mail, a baseline summary is provided, consisting of the first block of the text. A relation between e-mail features and summarization strategies is seen in Table 1, from more to less specific for e-mail domain.

4 Evaluation and Results

4.1 Establishing a golden standard

To tune and evaluate the performance of the system, a golden standard was 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[2]. 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 an indication of the content of the message. No guidelines were provided as to the length or type of the textual fragments to be marked, but relied on the communicative competence of the judges instead.

Since the intended goal of e-mail summarization is ill-defined, the golden standard served both as a representation of the goal and the reference ground to evaluate it. As a consequence of this double purpose, 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 *et al.* 01), who only used 8 e-mails.

4.2 Measures for evaluation

The ratio of agreement was used to assess the consistency of the gold standard, as it reflects the extent to which human judges agree on what makes a good summary.

Moreover, instead of the usual *precision* and *recall* metrics, the ratio of agreement was also used to compare human and automatic summaries, since it allows to equate automatic summaries to human summaries. This is specially adequate for tasks like summarization, where the notion of "the correct summary" is clearly prone to subjectivity. In effect, the golden standard can not be considered as the one and only "correct summary" for an e-mail, and therefore a peer-to-peer comparison seems more adequate.

Ratio of agreement between summaries, either human-human or automatic-human, was calculated at word level. The mean agreement between judges was 0.75, which indicates that human judges fairly agree on what makes a good summary. As a global measure of the system's

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

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

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

performance, we calculated how introducing the system as a human judge more affected the average ratio of agreement agreement.

To assess the performance of the system with respect to human judges, the ratio of agreement was calculated a second time, and manual summaries were randomly substituted by automatic summaries. If the system performed very differently from humans, the result of this second agreement would be much lower than the one between human judges alone. But, on the contrary, agreement values approached the ceiling established by humans: the obtained mean agreement was 0.66, thus decreasing only 0.1 with respect to human performance.

Additionally, content-based measures based in word overlap were used to account for equivalences in informativeness between human and automatic summaries, following the main trend in automatic summarization of e-mails (Mani 01). 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.

4.3 Discussion of results

Figure 3 presents an evaluation of CARPANTA summaries, performed on the 40-mail test corpus. Comparisons are grouped by the kind of strategy applied, so that it can be seen how well each strategy performs. Four different kinds of strategies can be distinguished: *baseline* (full mail, pyramidal, lead), *domain-specific* (subject, list, attachment, forward, appointment, question), *textual* (subjective, lexic, structural) and *combined* (textual, textual and documental). The summary chosen by CARPANTA is determined by the classification module, and is be the result of applying one of the previous strategies. There were no forwarded e-mails in the test corpus.

Ratio of agreement has been calculated between all automatic summaries provided for a given e-mail and every human summary available for that e-mail. Unigram and bigram overlap are also displayed. Additionally, Figure **??** displays some measures of robustness of the summarization strategies: coverage and compression rate. Coverage figures account for the percentage of mails in the corpus that can be summarized by each strategy, compression rate expresses the length of the summary as a ratio of the length

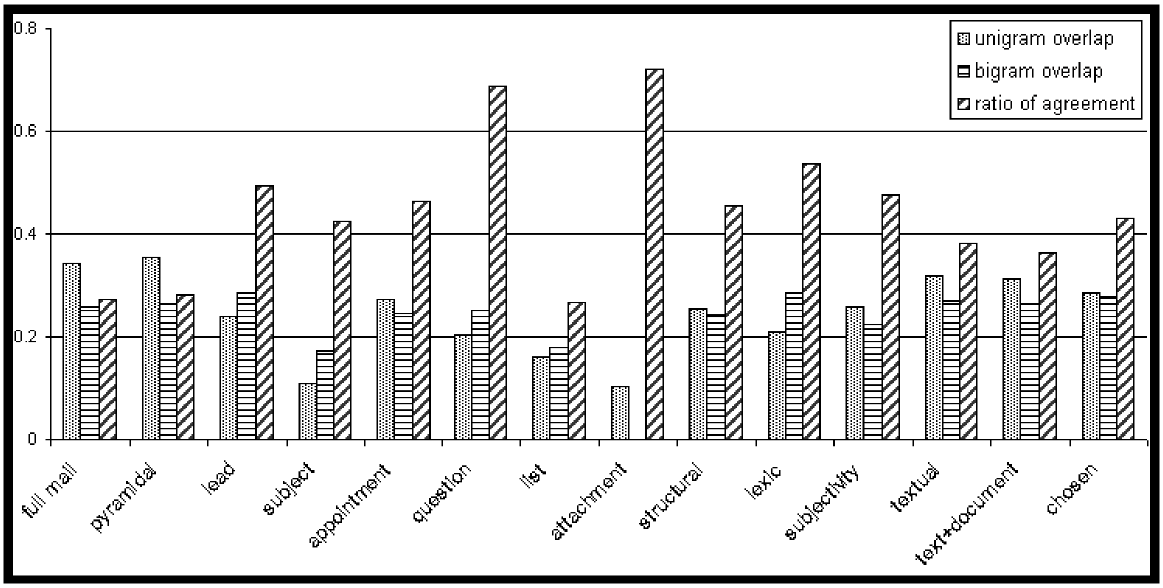

Figure 3: Average measures for the comparison of CARPANTA automatic summaries with human summaries. The content of automatic summaries is compared with each of their human counterparts by applying ratio of agreement, unigram and bigram overlap.

of the original message, so that longer summaries have a higher compression rate.

The *attachment* strategy has the highest average agreement values (almost 0.8), reaching an agreement with the golden standard at the level of human agreement. However, the coverage figure for this strategy is rather low, as is for most of the domain-specific strategies. In general, strategies with higher coverage present lower agreement values, and summaries exploiting e-mail specific knowledge show higher agreement with human judgement than textual ones, but the latter present a much higher coverage. A good trade-off between these two measures is provided by the strategies that combine more kinds of information, namely, *textual* and *textual and documental*.

However, very simple strategies, 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. More interestingly, providing the first sentence of the e-mail, the *lead*, gives even better results than the combined strategies, although its average informativity, measured by unigram overlap with the golden standard, is somewhat smaller: 23% overlap for *lead* against 31% for the combined.

The average agreement for the summary chosen by CARPANTA (0.63) is smaller than for other strategies, which indicates that an improvement on the classification of e-mails would improve the overall performance of the system.

4.4 Contribution to summary goodness

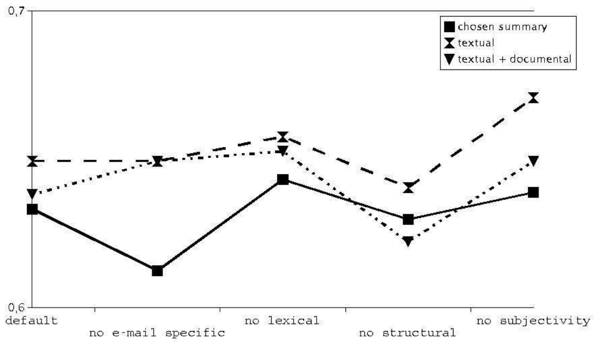

Figure 4: Mean agreement values for the *textual* and *textual and documental* strategies and for the summary chosen by CARPANTA, removing one kind of information at a time.

Figure 4 pictures the contribution of each kind of information to the goodness of the summaries within the three strategies that combine various kinds of information. For each of these strategies, the average ratio of agreement with the golden standard has been calculated for results with the default configuration of CARPANTA, and also ignoring each of the kinds of information that CARPANTA analyzes in e-mails.

In every case, agreement values range from 0.6 to 0.7, but it can be seen that ignoring domain-specific information deteriorates the quality of the chosen summary, while the other kinds of information introduce only minor changes in performance. For the *textual* and *textual and documental* strategies, it can be seen that structural information plays an important role in summary quality, and that subjectivity information has a negative effect, since ignoring it improves the resulting ratio of agreement.

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. The performance of the system has been evaluated with a corpus of human-made summaries, with high agreement with humans.

The contribution of various kinds of information to summary goodness has been studied, showing that domain-specific information yields high-quality summaries. This information will be incorporated to improve the accuracy of summarization strategies that merge heterogeneous information, as well as in the classification module.

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. Indeed, enhancements for Catalan and English are under development. Modules for automatic normalization and correction of input texts (Climent *et al.* 03) will also be included.

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 & Castellón 01) Laura Alonso and Irene Castellón. Towards a delimitation of discursive segment for natural language processing applications. In *First International Workshop on Semantics, Pragmatics and Rhetoric*, Donostia - San Sebastiàn, November 2001.

(Alonso *et al.* 00) A. Alonso, R. Folguerà, and C. Tebé. Del tecnolecte al sociolecte: consideracions sobre l'argot tècnic en català. *I Jornada sobre Comunicació Mediatitzada per Ordinador en Català (CMO-Cat)*, 2000. Universitat de Barcelona.

(Alonso *et al.* 02) Laura Alonso, Irene Castellón, and Lluís Padró. Design and implementation of a spanish discourse marker lexicon. In *SEPLN*, Valladolid, 2002.

(Atserias *et al.* 98) Jordi Atserias, Irene Castellón, and Montse Civit. Syntactic parsing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation*, Granada, 1998. LREC.

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

(Climent *et al.* 03) S. Climent, P. Gispert-Saüch, J. Moré, A. Oliver, M. Salvatierra, I. Sànchez, M. Taulé, and Ll. Vallmanya. Machine translation of newsgroups at the uoc. evaluation and settings for language control. *Journal of Computer-Mediated Communication*, 2003. in press.

(Fais & Ogura 01) L. Fais and K. Ogura. Discourse issues in the translation of japanese e-mail. *Proceedings of the Pacific Association for Computational Linguistics, PACLING 2001*, 2001.

(Ferrara *et al.* 90) K. Ferrara, H. Brunner, and G. Whittemore. Interactive written discourse as an emergent register. *Written Communication*, 8:8–34, 1990.

(Ginzburg & Sag 00) Jonathan Ginzburg and Ivan A. Sag. *Interrogative Investigations: the form, meaning and use of English Interrogatives.* Number Number 123 in CSLI Lecture Notes. CSLI Publications, Stanford, California, 2000.

(Herring 99) S. Herring. Interactional coherence in cmc. *Journal of Computer-Mediated Comunication*, 4(4), 1999. special issue on Persistent Conversation.

(Mani 01) Inderjeet Mani. *Automatic Summarization.* Nautral Language Processing. John Benjamins Publishing Company, 2001.

(Murray 00) D. E. Murray. Protean communication: the language of computer-mediated communication. *Tesol Quarterly*, 34(3):397–421, 2000.

(Sparck-Jones 99) Karen Sparck-Jones. Automatic summarising: factors and directions. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization.* MIT Press, 1999.

(Stede 03) Manfred Stede. Shallow - deep - robust. *Computerlinguistik - Was geht, was kommt? Computational Linguistics - Achievements and Perspectives*, 2003.

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

(Yates & Orlikowski 93) J.A. Yates and W.J. Orlikowski. 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, 1993. Center for Coordination Science Technical Report 150.