The Role of Domain-Specific Knowledge in Classifying the Language of E-negotiations

Mohak Shah, Marina Sokolova and Stan Szpakowicz
School of Information Technology and Engineering
University of Ottawa
Ottawa, ON, Canada
K1N 6N5
Email: {*mshah,sokolova, szpak*}*@site.uottawa.ca*

Abstract

Electronic negotiations are a special case of Web-based communication. Textual data collected during e-negotiations pose an interesting classification problem. Their distinct characteristics are a challenge for statistical Natural Language Processing and Machine Learning. We suggest a domain-specific text representation and argue that such a representation is necessary and sufficient to obtain a reliable classification. This also helps avoid the negative effect of non-standard text features common in data coming from electronic communication. To justify our claim, we report a variety of classification experiments that use our representation and contrast their results with the baseline and with other classification methods.

1 Introduction

New work patterns arise in businesses that adopt information technology on a large scale. Computer-mediated communication (CMC) is a standard term for human-to-human communication using computers. Examples of CMC are email, text-based chat and computer conferences [Climent *et al.*, 2003]. It brings new types of data for analysis. Text data gathered in CMC invite new applications of Machine Learning (ML) and statistical Natural Language Processing (NLP). Negotiations conducted through electronic means represent a rapidly growing category of CMC. We explore the data collected during bilateral electronic negotiations (e-negotiations) that last long enough to provide interesting material for our study [Kersten and Noronha, 1999].

Our ultimate goal is to study how the language used by the negotiators reflects the e-negotiation process, independent of any negotiation means. The longer an e-negotiation takes, the more elaborate the structure of the e-negotiation process becomes [Gebauer and Scharl, 1999]. Simpler e-negotiation may involve an exchange of well-structured business documents such as pre-defined contracts or retail transactions. A more complex process comprises numerous offers and counter-offers and has a high degree of uncertainty, which "results from the instability of the process environment, and from the unpredictability regarding the dynamic behavior of the organizational elements. The probability for changes of the situation and behavior as well as the extent to which they occur, play a central role." [Gebauer and Scharl, 1999].

The presence of electronic means poses an additional challenge. It has been observed [Herring, 2001] that when technology interferes with communication between humans, as in CMC, the participants' behaviour also undergoes various changes. Texts that appear in CMC have their peculiarities. [Murray, 2000] states that CMC uses "simplified registers" such as short sentences (**Go ahead.**), intended to make it easier for the reader to comprehend the message. [Climent *et al.*, 2003] have noted that texts of messages imitate human speech, using sound imitations (**Hm, Uh-ha**), letter repetitions (**sooon**), capitalization (**THANKS**). These special features can be extracted from the data and placed in a lexicon [Sokolova *et al.*, 2004]. Texts exchanged via CMC tend to be more syntactically informal [Yates and Orlikowski, 1993], highly erroneous and poorly edited [Climent *et al.*, 2003; Sokolova *et al.*, 2004]. These characteristics make the data challenging for the application of ML and NLP techniques. Another distinguishing property of text-based CMC is the limited nature of exchanges. There is no visual or acoustic information to help establish and strengthen personal contact or exhibit personal power, nor can the participants further their goals by resorting to sound or vision.

Our goal in this study is to obtain reliable classification of E-negotiations divided into two categories, namely completed (resulting in a successful compromise) and uncompleted (failing to reach a compromise). Briefly, we adopt the following method. We represent the text data in a way that captures the significant characteristics of the negotiation process independent of the electronic means – in this case a negotiation support system (NSS) – and does not bear the special features of CMC that tend to affects the learning adversely. This is a continuation of the work on language patterns in e-negotiations [Sokolova *et al.*, 2004]. We introduce a set of *domain-dependent* semantic categories and label our data with those categories.

We evaluate statistical characteristics of the data and build a semantic lexicon. Next, we find the semantic category to which the content words in the text (excluding stop words) belong more often than to others. We call this category *domain-specific*, and represent the e-negotiation data as bags of words built from this category. We classify data using various classifiers. Empirical results show that such a representation of the e-negotiation data provides stable outcomes for different classifiers. It also gives a marginally better outcome than a representation that uses non-textual NSS-dependent e-negotiation information [Kersten and Zhang, 2003]. We also favourably compare classification results with baseline, and use other data representations to justify that the data from the domain-specific category is necessary and mostly sufficient to obtain reliable results.

This work is part of an on-going research effort to better understand the patterns of e-negotiations in particular, and of CMC in general.

The paper is organized as follows. Section 2 describes the main characteristics of the *Inspire* system and the *Inspire* text data. Section 3 introduces the procedure. Experimental results appear in section 4 and previous results appear in section 5 followed by a brief analysis in section 6. In the last section we briefly discuss the limitations of the current results, state a few conclusions and suggest future work.

2 The *Inspire* Data

E-negotiation is a fast-growing Internet activity that includes exchange of email or other texts. In recent years the amount of data gathered through e-negotiation has achieved a volume that warrants applications of Data Mining (DM) and ML methods [Kersten and Zhang, 2003].

The largest collection of text data gathered in e-negotiation comes from the NSS *Inspire* [Kersten, 1999].

Inspire is a teaching and research tool widely used in university and college programs and on the Web. The language of negotiations is English, so all users must write in English, often their second language. There are no other restrictions on the users. *Inspire* supports users with a medium for conducting negotiations; it also provides the means of evaluating the negotiation process. The manuals and instructions about the negotiation process are posted on the Web. Each negotiation takes place between two people and should be completed in three weeks. Negotiation is completed if the virtual purchase has occurred within the designated time, and is uncompleted otherwise. The negotiators issue standard formal offers using the mechanisms supplied by *Inspire*, and may exchange free-form written messages. Messages either accompany offers or are sent between offers. In addition, the negotiators fill a pre-negotiation questionnaire, and may also fill a post-negotiation questionnaire. Below is an example of messages exchanged in an uncompleted negotiation:

Hallo Stacey, I hope you still enjoy our negotiation. I want to ask you if you would like to stay in contact with me after finishing the negotiation. we could exchange ideas about Australia, Germany, studies ... Best Regards Nadia My email-adresse: [...]

Dear Nadia, Please find enclosed the adjusted offer. I am aware of the financial urgency ass our company also deals with delivery and accounts holders. To ensure correct and prompt delivery of merchandise and asccociated funds, I have reconsidered the company's last offer. I hope that this offer is suitable and complies to your wishes. Best Regards Stacey

Dear Nadia, I would be more than pleased to keep in contact with you after our negitiation is completed. I have enjooyed it and i hope that i dont seem too forward or snobby. Yours Truely Stacey email: [...]

Dear Stacey, I think the offer would be acceptable. So let us do a deal on what we have discussed. Best Regards Nadia

The previous work on classifying the negotiation outcomes [Kersten and Zhang, 2003]dealt with the data extracted essentially from the three sources we listed, namely the pre- and post-negotiation questionnaires and the negotiation transcripts automatically generated by the NSS [Kersten and Noronha, 1999]. The questions formed the attributes for the data set (to be subsequently learned) and the responses to these questions were the attribute values. Some of these attributes are or may be confidential. Moreover, some other attributes depended on such factors as the sex of the negotiators, country of origin and so on. We call such attributes *strong*. There also are attributes whose values change in time and may depend on the circumstantial decisions during the process of negotiation; we call such attributes *dynamic* [Bazerman *et al.*, 2000; Gebauer and Scharl, 1999]. It is not advisable to assume that their value at any particular time can be used for learning.

The dynamic attributes are hard to quantify in advance. This would be true in most practical scenarios. For instance, *offers and ratings*, as well as *preference structures*, attributes used by [Kersten and Zhang, 2003], are generally dynamic in nature. There is a high probability that they change values over time and during negotiations. Other factors might affect such attributes. Here is a possible scenario: an offer is initially unacceptable to the buyer; during the negotiations the buyer may accept this offer when it comes as part of some package deal that tends to be a compromise acceptable to both parties. Such situations might strongly affect the dynamic attributes [Drake, 2001], so they must be handled accordingly. In essence, it is not justified to use the values of dynamic attributes at any time. The way in which [Kersten and Zhang, 2003] deal with them can be called the static use of dynamic attributes.

As a consequence, while we extract the data from which reliable learning is possible, we must address all these issues. Learning should not rely heavily on *strong* attributes, but the extracted data should still reflect the characteristics of the data associated with the class of e-negotiations. The *DSDR* procedure that we describe in the next section attempts to address these issues. Also, using the dynamic attributes statically [Kersten and Zhang, 2003] is not justified. In the case of data representation using the *DSDR* procedure, we avoid the use of

dynamic attributes and hence their inherent complexity for classification.

The *Inspire* text data available to us consists of the transcripts of 2557 negotiations, 1427 of them completed. Each negotiation involves two people, and one person participates in only one negotiation. The number of the data contributors is over 5000. The data contains 1,514,623 word tokens and 27,055 word types. The data bear all the typical characteristics of CMC as discussed in Section 1, including high volume of personal information. In addition to business discussions, negotiators also discussed their studies, hobbies, personal affairs, and so on.

3 Representing Domain-Specific Data

We now describe a procedure that we call **D**omain-**S**pecific **D**ata **R**epresentation (*DSDR*). We propose to use the *domain-specific* words to represent the data, and we suggest that such words are important for the classification and prediction. We will explain what we call the *Domain-Specific* categories of the data as we describe the procedure.

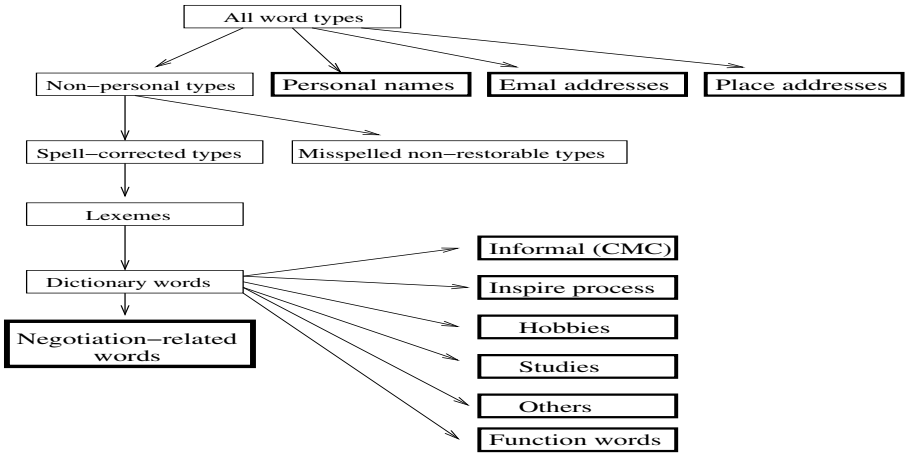
We first construct a unigram model from the original text data. The model yields a set of unigrams (different word types) and their number of occurrences in the text. Next, we preprocess the list of unigrams to remove stop words (those are mainly function words). Other operations might be performed – spelling corrections, stemming, lemmatization – but the characteristics of the data do not justify their application. Such operations might in fact adversely affect the results. For example, the word *message* is used in the data in the regular sense and does not indicate a positive or negative development in negotiations. On the other hand, the word *messages*, which would be stemmed or lemmatized to the word *message*, is used in 80% of the situations when trouble in communication is detected, so it indicates a negative development. Such observations suggest that the data in its original form can be much more helpful.

Although we would like to work mainly with the unigram model, we also create bigram and trigram models of the data in order to perform the *domain-specific* word sense disambiguation of various unigrams. Here is an example of how this approach helps. Two most frequent bigrams that include the word *policy* are *return policy* and *returns policy*, which cover about 66.5% occurrences of the word *policy*. *Returns* is one of the four issues negotiated in *Inspire*'s standard problem. We therefore tag the word *policy* as a negotiation-related word.

It should be noted that due to the small size of our documents, we cannot use the standard information retrieval or data mining techniques to model our data, but the Good-Turing model fits our data well with Katz smoothing. Refer to [Shah *et al.*, 2004] for more details on empirical results of the statistical modelling of e-negotiations. Katz smoothing results in a relatively low cross-entropy for our model, so we use the cut-off suggested by this model to smoothen our data. We remove all the unigrams with occurrence counts below 6. The removal of such data also serves to remove the personal information from the data, as well as the rare words which might not be statistically representative of the data.

We did not investigate in depth the distribution of personal information. Our general study of the data, however, suggests that the presence of a personal email address is a trustworthy indicator of the personal nature of (part of) an Inspire message. Email addresses are usually exchanged when the partners perform self-disclosure. We extracted 512 negotiations that contained personal email addresses and tested the distribution of the occurrences of unigrams corresponding to personal information. 90% of such unigrams had less that 6 occurrences.

Figure 1: Procedure of defining semantic categories



Now, we automatically build a semantic lexicon and tag the remaining words with semantic tags. We apply `ispell` for spelling corrections and LDOCE [Summers, 2003] to semantically tag the data. The DSDR procedure is briefly described in Figure 1. For a complete description on the tagging procedure, see [Sokolova *et al.*, 2004].

We analyze semantic categories of 200 most frequent unigrams and correct them manually. We use the following semantic categories to tag the word types: **Negotiation-related**, **Studies**, **Informal (CMC) words**, ***Inspire* process**, **Hobbies**, **Personal Names**, **Email addresses**, **Place addresses**, **Function words**, **Others**. The words in the *Negotiation-related* category are those to which we refer as *domain-specific* data, since they are quite specific to e-negotiations.

The *Inspire* corpus that we deal with has a different distribution of unigrams than the well-known corpora: *Brown Corpus Manual* [Francis and Kucera, 1964] and *Wall Street Journal* (WSJ). In our case, the negotiation-related words rank higher. For instance, the word *offer* appears among the 10 most frequent words in our corpus; only function words appear among the top 10 frequent words in Brown and WSJ.

Once a semantic lexicon has been built, we calculate the percentage of occurrence among the 100 most frequent unigrams of the words of each semantic category except function words. The largest percentage comes from the words of the negotiation-related category – as expected (see Table 1).

We now consider only the negotiation-related words to further build our data set and to perform learning and classification. We give further experimental details in next section.

4 Experimental Results

We must experimentally verify our claim that the e-negotiation data can be represented by a subset of domain-specific unigrams and can be classified relatively accurately. To do this, we form a bag of *Negotiation-related*

Table 1: Distribution of category types (excluding function words) in 100 most frequent unigrams

Category	% **of Words**
Negotiation-related	57.9
Studies	0
Informal (CMC) words	0
Inspire process	5.4
Hobbies	0
Personal Names	0
Email addresses	0
Place addresses	0
Others	36.7

Table 2: The accuracy on e-negotiations represented with *DSDR*

Classifier	**Accuracy**(%)
BL	55.8
NB	60.62
IBK	70.6
SMO	71.72
DT	72.39
DS	73.13
c5.0	**75.4**

Table 3: Classification of positive negotiations (percentage).

Classifier	**Precision**	**Recall**	**F-measure**
BL	55.8	100	71.6
NB	53	80.2	64.3
IBK	**76**	69	72
SMO	75.8	72.5	74
DT	71.2	87.2	78.4
DS	71.4	85.6	77.8
c5.0	73.3	**87.7**	**79.9**

words that we identified in section 3. This gives us a dataset of dimensionality 123. We add to this a count of the number of unigrams in each example that do not belong to the *Negotiation-related* semantic category. So, for each example we have bags of words with 124 attributes. Each of the first 123 attributes in each example represents the number of occurrences of the corresponding unigram from our domain-specific (*Negotiation-related*) category while the last attribute gives the total number of other unigrams present in the example. Each example is labeled positive if the corresponding negotiation resulted in a completion, and negative otherwise.

We have a total of 2557 examples in our data set, of which 1427 are positive and 1130 negative. We report the average tenfold cross-validation results for all the experiments. These results were obtained over the set of parameters for each classifier that yielded the highest classification accuracy. The parameters were selected using an exhaustive search in the space of possible parameters. We have employed several classifiers freely available in the Weka suite [Witten and Frank, 2000]: Instance-based using 20-nearest neighbor (IBK), Naive Bayes classifier (NB), Decision Stumps (DS), Decision Tables (DT) and linear SVM (SMO). For decision trees we have used C5.0. BL indicates the Baseline for our data set.

The accuracy results appear in Table 2. We present the best accuracy achieved by each classifier after we have performed exhaustive search on adjustable parameters. Precision, recall and F-measure are calculated with respect to the completed negotiations and reported in Table 3.

Now, in order to verify our claim that the data representation using only the domain-specific words is necessary

Table 4: The accuracy e-negotiations represented with 500 top words

Classifier	**Accuracy**(%)
BL	55.8
IBK	65.8
SMO	N/A
NB	63.4
DT	N/A
DS	73.19
c5.0	**75.51**

Table 5: Classification of positive negotiations on top 500 words (percentage).

Classifier	**Precision**	**Recall**	**F-measure**
BL	55.8	100	71.6
IBK	67.78	60	63.75
NB	71.2	46.4	55.83
DS	68.32	82.97	74.75
c5.0	73.39	88	79.8

Table 6: The accuracy on e-negotiations represented with 123 mixed words

Classifier	**Accuracy**(%)
BL	55.8
IBK	67
SMO	69.11
NB	58
DT	70.87
DS	72.91
c5.0	74.2

Table 7: Classification of positive negotiations on top 123 mixed words (percentage).

Classifier	**Precision**	**Recall**	**F-measure**
BL	55.8	100	71.6
IBK	64.8	73.3	68.2
SMO	68.4	68.8	68.6
DT	65.03	86.54	74.19
NB	71.97	41.56	52.25
DS	68.3	81.08	73.73
c5.0	67.78	87.04	76.18

and sufficient, we perform the following two sets of experiments. First we represent the data using the 500 most frequent unigrams. That is, we form bags of words for each negotiation using the number of occurrences of the 500 most frequent unigrams over the whole set of negotiations. We perform tenfold cross-validation training and report the corresponding results in Tables 4 and 5. N/A means that the corresponding classifier were running at least 5 times more slowly than they did with our DSDR representation of data.

In the second set of experiments, we represent the data using a mixed set of 123 unigrams (excluding function words). We define a mixed set of unigrams as one that has a subset of unigrams belonging to the *Negotiation-related* category and a subset of other randomly chosen unigrams with at least 6 occurrences. We again form bags of words for each negotiation using the number of occurrences of each of 123 unigrams thus selected, and perform tenfold cross validation over the classifiers. The results of these experiments are reported in Tables 6 and 7.

5 Previous work on the classification of E-negotiations

Previous studies on classifying e-negotiations did not consider the language aspect of negotiations. Working with *Inspire* data, [Kersten and Zhang, 2003] used data mining to classify 1525 negotiations as success or failure based on various factors including the characteristics of the negotiations. Each negotiation was represented by the number of offers sent, regularity with which they were sent, time when they were sent, with special attention paid to the time of the last offer, and so on. Also, this research, as we said earlier, used some dynamic attributes

Table 8: (Best) accuracy results from non-textual classification on 1525 negotiations

Classifier	**Accuracy** (%)
Neural Networks	59.28
Loglinear Regression	62.4
Decision Trees	75.33

statically. The results are presented in Table 8. The precision, recall and F-measure values are not available. The details of these experiments appear in [Kersten and Zhang, 2003]. Some preliminary work applied NLP methods to the preprocessing of data, and the building of a semantic and syntactic lexicon. The classification results have been reported in [Sokolova *et al.*, 2004]. The results in our case clearly show that a relatively comparable (in fact marginally better) accuracy can be obtained when only domain-specific knowledge is used to represent the data. They also suggest that language is important for the outcome of negotiations.

6 Analysis of the Experimental Results

We have observed, as discussed in section 3, that among the most frequent unigrams (word types) there are about 58% of negotiation-related words. It is obvious that we get the similar classification accuracy as when using the most frequent 500 unigrams (Tables 4 and 5), by using the DSDR representation of data (Tables 2 and 3). However, the DSDR representation is succinct and makes use of relevant domain-specific knowledge. This serves to show that our representation retains the knowledge about the data and hence is sufficient to obtain a reliable classification.

On the other hand, we see that using a mixed set of 123 unigrams (from among the first 500 unigrams) gives marginally worse results (Tables 6 and 7). Also, it should be noted here that the overlap between these 123 mixed unigrams and the 123 unigrams chosen according to the DSDR procedure is approximately 50%. This indeed suggests that the domain-specific knowledge is necessary for better classification.

We also see that our results using the DSDR representation are better than those produced by [Kersten and Zhang, 2003] using the non-textual representation of data (Table 8). Moreover, our data representation makes use of the relevant domain-specific knowledge and is independent of dynamic attributes. The results of [Kersten and Zhang, 2003] rely on such dynamic attributes, but statically; this seems unjustified.

The results show that the accuracy of classification of uncompleted negotiations is lower than the accuracy of classification of completed negotiations, with the exeption of Naive Bayes classification. We have looked for reasons which lead to the difference in the accuracy results. Analyzing the preformance results of Naive Bayes we conlude that the assumption of conditional independence of features, i.e. negotiation-related words, **is not met** in completed negotiations and **is met** in uncompleted negotiations. We conclude that the negotiation-related words are correlated in completed negotiations and are not correlated in uncompleted negotiations. We partially attribute the difference in accuracy to inaccurate system's labelling. The labels given by the Inspire system do not always correspond to the real outcome of negotiations: negotiation is labelled as completed if the box "Accept" is checked, in all other cases negotiation is labelled as uncompleted. However, analysis of the data has shown that 3-5 % of uncompleted negotiations were finished with the agreement of participants to

accept an offer.

7 Conclusion

Continuing the study of the language patterns of e-negotiations [Sokolova *et al.*, 2004], we propose a new representation procedure for the e-negotiation text data. The representation, which we call DSDR, captures the relevant characteristics of such data while leaving out the adverse CMC traits. The empirical results show that such a representation of the e-negotiations provides stable outcomes for different classifiers and gives a marginally better outcome than classification using non-textual e-negotiation information [Kersten and Zhang, 2003]. We also show that the domain-specific knowledge is necessary and sufficient for reliable classification. The approach has another important aspect in terms of the dependence on the NSS that collects the data. The results of [Kersten and Zhang, 2003] crucially depend on the NSS *Inspire*. Our results do not rely on any such NSS. They only require the availability of verifiable domain-specific knowledge and texts that accompany e-negotiations.

The DSDR procedure described in this paper is generic, so it can prove useful in any typical Web-based communication scenario. Further research is needed, however, to confirm the applicability of the procedure in other domains. Our current work is a step in realizing the importance of such domain-specific knowledge and its use for practical learning tasks.

Acknowledgment

This work is partially supported by the Social Sciences and Humanities Research Council of Canada and by the Natural Sciences and Engineering Research Council of Canada.

References

[Bazerman *et al.*, 2000] M. H. Bazerman, J. R. Curhan, D. A. Moore and K. L. Valley. Negotiation. *Annual Review of Psychology*. http://www.findarticles.com/cf0, 2000.

[Chen and Goodman , 1998] S. F. Chen, J. Goodman. *An Empirical Study of Smoothing Techniques for Language Modeling*. Tech. Report TR -10-98, Center for Research in Computing Technology, Harvard University, Cambridge, Massachusetts, 1998.

[Climent *et al.*, 2003] S. Climent, J. Mor, A. Oliver, M. Salvatierra, I. Snchez, M. Taul and L. Vallmanya. Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation *Journal of Computer-Mediated Communication*[On-line], 9(1), 2003. http://www.ascusc.org/jcmc/vol9/.

[Drake, 2001] L. E. Drake. The Culture-Negotiation link. *Human Communication Research*, **27**, 3, 317–349, 2001.

[Francis and Kucera, 1964] W. M. Francis and H. Kucera. *Brown Corpus Manual of Information*. Department of Linguistics, Brown University, 1964. http://helmer.aksis.uib.no/icame/brown/bcm.html

[Gebauer and Scharl, 1999] J. Gebauer, A. Scharl Between flexibility and automation: An evaluation of Web technology from a business process perspective. *Journal of Computer-Mediated Communication* [On-line], 5(2), 1999. http://www.ascusc.org/jcmc/vol5/

[Herring, 2001] S. C. Herring. Computer-mediated discourse. In D. Tannen, D. Schiffin, H. Hamilton (eds.), *Handbook of discourse analysis*, 612–634, Oxford, Blackwell, 2001.

[Jurafsky and Martin, 2000] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, 2000.

[Katz, 1987] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP 35(3), 400-401, March, 1987.

[Kersten, 1999] G. E. Kersten. *The Science and Engineering of E-negotiation: An Introduction*. InterNeg Report 02/03, 2003. interneg.org/interneg/research/papers/

[Kersten and Noronha, 1999] G. E. Kersten and S. J. Noronha. WWW-based Negotiation Support: Design, Implementation, and Use. *Decision Support Systems*,25, 135–154, 1999.

[Kersten and Zhang, 2003] G. E. Kersten and G. Zhang. Mining Inspire Data for the Determinants of Successful Internet Negotiations. *Central European Journal of Operational Research*, 11(3), 297-316, 2003.

[Murray, 2000] D. E. Murray. Protean communication: The language of computer-mediated communication. *TESOL Quarterly*, 34(3), 397–421, 2000.

[Paul and Baker, 1992] D. B. Paul and J. M. Baker. The Design for the Wall Street Journal-based CSR Corpus. *Proc ICSLP'92*, Kobe, 1992.

[Rosenfeld, 2000] R. Rosenfeld. Two Decades of Statistical Language Modeling: Where Do We Go from There?. *Proceedings of IEEE*, 88(8), 1270–1278, 2000.

[Shah *et al.*, 2004] M. Shah, M. Sokolova and S. Szpakowicz. *Using Domain-Specific Knowledge to Classify E-negotiations*. InterNeg Working Paper, INR 07/04, 2004. http://interneg.org/interneg/research/papers/index.html.

[Sokolova *et al.*, 2004] M. Sokolova, S. Szpakowicz and V. Nastase. *Automatically Building a Lexicon from Raw Noisy Data in a Closed Domain*. InterNeg Working Paper 01/04, 2004. http://interneg.org/interneg/research/papers/index.html.

[Sokolova *et al.*, 2004] M. Sokolova, S. Szpakowicz and V. Nastase. Using Language to Determine Success in Negotiations: A Preliminary Study. *Advances in Artificial Intelligence 17*, pp 449-453, Springer, 2004.

[Summers, 2003] D. Summers (ed.) *Longman Dictionary of Contemporary English*, Fourth edition. Pearson Education. Longman, 2003.

[Witten and Frank, 2000] I. Witten, E. Frank. *Data Mining*, Morgan Kaufmann, 2000. http://www.cs.waikato.ac.nz/ml/weka/

[Yates and Orlikowski, 1993] J. A. Yates, W. J. Orlikowski. *Knee-jerk anti-LOOPism and other e-mail phenomena: Oral, written, and electronic patterns in computer-mediated communication*. MIT Sloan School Working Paper 3578-93, Center for Coordination Science Technical Report 150, 1993. http://ccs.mit.edu/papers/CCSWP150.html