InterNeg Research Papers INR 07/04

Using Domain-Specific Knowledge to Classify E-negotiations

Mohak Shah
SITE, University of Ottawa,Canada

Marina Sokolova
SITE, University of Ottawa,Canada

Stan Szpakowicz
SITE, University of Ottawa,Canada
{mshah,sokolova,szpak}@site.uottawa.ca

June 18, 2004

Abstract

Texts exchanged in business-related Computer-Mediated Communication, or CMC, differ from texts exchanged in other business situations. CMC data have a high concentration of non-standard textual features. The fast-growing amount of business CMC data offers opportunities for the application of statistical Natural Language Processing and Machine Learning methods, especially for text-classification purposes. We suggest a domain-specific text representation that helps avoid the negative effect of the non-standard text features. We report a variety of classification results that use this representation. We also build a statistical language model for the data and present its results. This is the first study on both statistical modeling of such data and their classification solely through text representation.

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. CMC includes email, text-based chat, computer conferences, and so on [4]. 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 type of CMC. We explore the data that come from bilateral electronic negotiations (e-negotiations) that last long enough to provide interesting material for our study [11].

Our 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 [6]. 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."[6].

The presence of electronic means poses an additional challenge. It has been observed [7] 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. [13] 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.* [4] have noted that texts of messages imitate human speech, using sound imitations **Hm, Uh-ha** and dots, letter repetitions **sooon**, capitalization **THANKS**. These special features can be extracted from the data and placed in a lexicon [15]. Texts exchanged via CMC tend to be more syntactically informal [18], highly erroneous and poorly edited [4, 15]. These characteristics make for challenging data for 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.

In the present study we aim to 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 [16]. 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 words in the text belong more often than to others, omitting stop-words. 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 [12]. We also build a statistical model of the data and compare the cross-entropy obtained to the cross-entropy of English texts. This, to the best of our knowledge, is the first statistical model of e-negotiation text data and of the text data that come from bilateral CMC.

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. A statistical model that fits the data is explained in Section 4. Experimental results appear in section 5. 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 [12].

The largest collection of text data gathered in e-negotiation comes from the NSS *Inspire* [10]. *Inspire* is a teaching and research tool widely used in university and college programs and on the Web. The language of negotiations is English, hence all users must speak English, often as a 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.

The previous work on classifying the negotiation outcomes [12], as well as other studies, were performed on 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 [11]. 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. 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 [1, 6]. 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*, *preference structures* are two such attributes used by [12] that are generally dynamic in nature. There is a high probability that these attributes change their values over time and over the course of negotiations. Other factors might affect such attributes. Here is a possible scenario. At the beginning of the negotiation, an offer is unacceptable to the buyer; during the course of 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 [5].

These characteristics of the data suggest that learning from such data, and predictions based on them, ought not to depend on such strong attributes. However, the 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 aims at targeting these issues.

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). There are other operations that might be performed – spelling corrections, stemming, lemmatization – but the characteristics of the data do not justify their application. These 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 show 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 such an 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 that are 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. In the next section, we describe the Good-Turing model and compare two smoothing techniques, Katz smoothing and Kneser-Ney smoothing. 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.

Now, we automatically build a semantic lexicon and tag the remaining words with semantic tags. We apply `ispell` and LDOCE [15]. 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
- Email addresses
- Place addresses
- Function words
- Others

The words in the *Negotiation-related* category are those that we refer to as *domain-specific* data, since they are quite specific to e-negotiations.

Table 1: Ranks of negotiation-related words

Type	**Inspire rank**	**Brown rank**
offer	8	1320
price	20	5000+
delivery	27	4993
accept	35	5000+
days	40	229
payment	43	2053
company	49	5000+
quality	55	908
negotiation	57	5000+
returns	64	3166

Table 2: 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
Email addresses	0
Place addresses	0
Others	36.7

The *Inspire* corpus that we deal with has a different distribution of unigrams than the well-known *Brown* corpus [2] and *Wall Street Journal* corpus. 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 the two widely used corpora we have named. In Table 1 we compare the ranks of 10 most frequent negotiation-related words from the *Inspire* data with their ranks in the *Brown* corpus. 5000+ means that the word's rank exceeds 5000.

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 2). 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 Section 5. Before proceeding to the experimental results, we give a brief account of the statistical model for our data.

4 Data Modeling

The *Inspire* data show many interesting characteristics with regard to statistical modeling, quite different than the characteristics of such widely used NLP corpora as *Brown* or the *Wall Street Journal* corpus. Unlike those standard corpora, ours has a higher token-type ratio, regular percentage of rare words [8] and high percentage of most frequent words. There also seem to exist no references to work on modeling text data from bilateral CMC. We present here a preliminary study of statistical modeling of such data. N-gram models are arguably the most widely used models for statistical and language modeling purposes. An important concern in such modeling is *smoothing* that helps incorporate the knowledge of the previously unseen N-grams. We use Katz smoothing [9], for a few theoretical and practical reasons. Katz smoothing generally performs well on data with a high token-type ratio. It is easier to implement. It has very few parameters and does not require a validation set[3, 8] for adjustable parameters. In our future work, however, we want to incorporate some data-dependent information too, for example, the negotiation specific distinction between buyers and sellers. The size of our data is another argument for Katz smoothing.

We now briefly describe Katz smoothing and show the cross-entropy results for it and for another attractive technique, Kneser-Ney smoothing. The latter perform well with respect to collocations; this might be helpful since we work in a closed domain with a fixed topic.

Katz smoothing basically combines the higher-order models with lower-order models, extending the Good-Turing estimate. For simplicity, we consider the case of a *bigram* model. Katz smoothing for n-gram models of higher orders is analogous. In fact, any Katz n-gram model is defined in terms of the Katz $(n-1)$-gram model. A sentence s is composed of words $w_1 ... w_l$. The corrected count of a bigram w_{i-1}^i with count $r = c(w_{i-1}^i)$ is given by:

$$c_{katz}(w_{i-1}^i) = \begin{cases} d_r r & if \quad r > 0 \\ \alpha(w_{i-1}) p_{ML}(w_i) & if \quad r = 0 \end{cases}$$

We discount all the bigrams with non-zero count with the Good-Turing discount ratio d_r which is approximately $\frac{r^*}{r}$ where r^* is the Good-Turing estimate of r. The counts thus subtracted are distributed among the zero-count bigrams according to the unigram model distribution, that is, the next lower-order distribution. The value of $\alpha(w_{i-1}^i)$ is chosen with the constraint $\Sigma_{w_i} c_{katz}(w_{i-1}^i) = \Sigma_{w_i} c(w_{i-1}^i)$. The value of $\alpha(w_{i-1})$ is given by:

$$\alpha(w_{i-1}) = \frac{1 - \Sigma_{w_i : c(w_{i-1}^i) > 0} p_{katz}(w_i | w_{i-1})}{\Sigma_{w_i : c(w_{i-1}^i) = 0} p_{ML}(w_i)} = \frac{1 - \Sigma_{w_i : c(w_{i-1}^i) > 0} p_{katz}(w_i | w_{i-1})}{1 - \Sigma_{w_i : c(w_{i-1}^i) > 0} p_{ML}(w_i)}$$

Now, normalizing yields the value for $p_{katz}(w_i | w_{i-1})$:

$$p_{katz}(w_i | w_{i-1}) = \frac{c_{katz}(w_{i-1}^i)}{\Sigma_{w_i} c_{katz}(w_{i-1}^i)}$$

Large counts are generally considered reliable and hence d_r is taken as 1 for any count $r > k$ where k is suggested by Katz to be 5. For lower counts $r \leq k$, d_r is found according to the following equation:

$$d_r = \frac{\frac{r_*}{r} - \frac{(k+1) n_{k+1}}{n_1}}{1 - \frac{(k+1) n_{k+1}}{n_1}}$$

Finally, the Katz unigram model is taken to be the Maximum Likelihood model. For more details on Katz smoothing see[9, 3].

Table 3: Cross-entropy results

Data	GTK model	KN model
partial	5.69	5.75
whole	5.66	N/A

Now, we need to evaluate the goodness of the fit for this model on our data. The standard measure for evaluating statistical models is cross-entropy

$$-\frac{1}{n}\sum_{i=0}^{n} log(P(w_i))$$

where n is the number of words in the test set, $P(w_i)$ is the probability of the appearance of the word w_i in the test data. Low cross-entropy means that the data is predictable, higher cross-entropy indicates high uncertainty in the data. The cross-entropy of English texts ranges from around 5.64 to 9.70, depending on the type of text [3].

A model with the lower cross-entropy on the test set models the data better [3]. We chose to use trigram models with the modified Kneser-Ney smoothing method (KN model) [3] and the Katz variant of Good-Turing smoothing method, with $k = 5$ (GTK model), where k is the number of occurrences of a unigram in the data [3].

In the first step of building the models we used part of the *Inspire* data, with 648,931 tokens of which 581,631 were in the training set. In the second step we used all our *Inspire* data, 1,107,447 tokens for training and 398,703 for testing (other tokens were filtered out as non-English words). On the partial data, the cross-entropy obtained by the KN model was higher than the cross-entropy obtained by the GTK model. The KN model ran significantly longer than the GTK model. We therefore built only the GTK model on the whole data. The cross-entropy results are reported in Table 3. The results show that the *Inspire* data are highly predictable.

5 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* 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 10-fold cross-validation results. We have employed several classifiers freely available in the Weka suite [17]: Instance-based using 20-nearest neighbor (IBK), 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 4. 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 5.

Table 4: The accuracy of various classifiers on 2557 e-negotiations represented using the *DSDR* procedure

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

Table 5: Classification of positive negotiations (percentage).

Classifier	**Precision**	**Recall**	**F-measure**
BL	55.8	100	71.6
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**

Previous studies on classifying e-negotiations did not consider the language aspect of negotiations. Working with *Inspire* data, Kersten and Zhang [12] 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 offers were sent, time when the offers were sent, with special attention paid to the time of the last offer, and so on. Their results are presented in Table 6. The precision, recall and F-measure values are not available. The details of these experiments appear in [12].

Another approach applied to the classification of such data used NLP methods to preprocess data, and to build a semantic and syntactic lexicon. The results have been reported in [15]. 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 the language aspects are important to the outcome of negotiations.

Table 6: (Best) accuracy results from non-textual classification on 1525 negotiations

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

6 Conclusion and Future Work

Continuing the study of the language patterns of e-negotiations [16], we propose an approach to represent CMC text data in a way that captures the chief characteristics of such data while leaving out the adverse CMC traits. The empirical results show that such a representation of the CMC data (data from e-negotiations in our case) provides stable outcomes for different classifiers and gives a marginally better outcome than classification using non-textual e-negotiation information [12]. The approach has another important aspect in terms of the dependence on the NSS that collects the data. The results of [12] crucially depend on the NSS *Inspire*. Our results do not rely of any such NSS. They only depend on the availability of verifiable domain-specific knowledge and its language aspects.

Even though the DSDR procedure described in this paper is a generic approach and can prove useful in any typical CMC scenario, considerable research is needed to confirm this. Work is also necessary in the area of fully automating the procedure. For instance, the tools like `ispell` and LDOCE [15] are not specifically suitable for such *domain-specific* tasks and need manual intervention. There are also some areas that can be improved to compensate for the loss of *domain-specific* information due to the characteristics of the CMC data. It should be noted that there is still plenty of room for the loss of such information, for example, when such information is abbreviated. Adequate resources are not yet available to help take care of such issues, especially in the domain of e-negotiation. Also, the *Inspire* data set suffers from problems that arise from the inherent dependency on the NSS, such as mislabelled examples. These problems will be part of our future work in this direction; we might apply supervised and/or unsupervised learning to solve them. Our current work, however, is a step forward 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

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

[2] Brown Corpus Manual. http://helmer.aksis.uib.no/icame/brown/bcm.html

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

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

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

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

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

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

[9] S. M. Katz. 1987 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.

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

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

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

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

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

[15] M. Sokolova, S. Szpakowicz, and V. Nastase 2004 *Language Patterns in Text Messages of Electronic Negotiations: A preliminary Study* InterNeg Report 05/04 http://interneg.org/interneg/research/papers/

[16] M. Sokolova, S. Szpakowicz, and V. Nastase 2004 *Using Language to Determine Success in Negotiations: A Preliminary Study* Proc 17th Conference of the Canadian Society for Computational Studies of Intelligence pp 449 -453

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

[18] J. A. Yates, W. J. Orlikowski. 1993 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. http://ccs.mit.edu/papers/CCSWP150.html