Feature Selection for Electronic Negotiation Texts

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

Abstract

Various feature selection and data representation methods have been proposed for text data collected from electronic negotiations. We compare two broad classes: process-based and corpus-based feature selection. In particular, we study the informativeness and representativeness of each method from these classes with respect to the classification of outcomes of electronic negotiations. Our empirical results are a quantitative basis for our analysis.

1 Introduction

Texts exchanged in electronic negotiations (e-negotiations) contain signals that may indicate the successful or unsuccessful outcome. In order to extract such signals, it is essential to find an effective feature selection method and a suitable data representation to enable learning in this environment. Various methods to address this issue, with various biases, have been proposed. In this paper, we introduce two broad classes of such methods, with important commonalities. We further analyze the methods in each class, looking for those that result in an optimum feature subset and data representation for texts coming from e-negotiations. We focus on identifying the learning settings that better assist the prediction of negotiation outcomes. The important components of such settings are features, their representation and the learning paradigm. The quality of the classification of the negotiation outcomes is one of the evaluation measures. Note that although we reduce the classification of negotiation outcomes to the classification of negotiation texts, our procedure differs from standard text classification. For an overview of machine learning methods and their application to text classification, including different types of features refer to (Sebastiani, 2002).

The first class that we consider contains the methods that exploit the knowledge of the negotiation process and the strategies employed when two parties negotiate. The former, based on the identification of negotiation-related words, was introduced in (Shah et al, 2004). The latter, using strategy-related features, was introduced in (Sokolova and Szpakowicz, 2005). The data representation based on negotiation-related features benefits from the knowledge of the negotiation. On the other hand, the strategy-related data representation relies on more general knowledge of the influence strategies that negotiators employ to reach a beneficial agreement. However, both these methods rely on the knowledge of the *process* of negotiation, though at different levels. Hence, we place them together under the umbrella of *process-based* data representation.

The second class that we discuss here contains methods that identify a representative subset of features by considering the statistical characteristics of the data under investigation. One such method, quite popular, represents data with the most frequent n-grams; often it is a unigram representation ($n = 1$). We also introduce an approach that relies on features whose frequency behaviour varies between data classes. Those are features that occur more frequently in one class than in the other (for example, in successful rather than unsuccessful negotiations). All these methods work with corpus statistics; we name them collectively *corpus-based* data representation.

Having defined the classes of data representation and feature selection methods for e-negotiations, we continue our analysis to address two issues:

- which set of features is better suited to represent e-negotiation texts so as to classify them on negotiation outcomes;

- which representation gives better insights into the negotiations themselves.

We employ various learning paradigms to examine the behaviour and usefulness of each representation.

In addition, we also examine whether the presence of selected features is important or the frequency of occurrence matters equally to each candidate feature selection method. Finally, we show that the *process-based* approach fares better in terms of the classification accuracy than the *corpus-based* approach. The correct identification of successful and unsuccessful negotiations increases when the feature sets result from process-based approaches. Pinpointing the most representative features should help predict the negotiation outcome better during the negotiation itself, and warn the negotiators when their language use may lead to a failure.

The insights gained will be useful in studying and extracting knowledge about specific negotiation problems such as strategies, tactics, negotiation moves and ways in which negotiation partners exert influence on each other and in identifying the appropriate feature sets for such tasks.

The remainder of the paper is organized as follows: Section 2 introduces the environment of e-negotiations and the specifics of the e-negotiation data. Section 3 describes the feature selection methods that we investigate; they all come from the two broad classes discussed earlier in the paper. Section 4 discusses in detail the experimental setting and reports the classification results. Section 5 presents an analysis of, and insights into, the behaviour and usefulness of various methods in the light of our experiments. Finally, Section 6 highlights the main findings and future directions.

2 E-negotiation Data

E-negotiations occur in various domains (for example, labour or business) and involve various users (for example, negotiators or facilitators). As in traditional negotiations, e-negotiation participants have established goals and exhibit strategic behaviour (Brett, 2001). The negotiation outcome (success or failure) results from these strategic choices. E-negotiations held by humans, however, share the uncertainty intrinsic to any human behaviour.

Text messages exchanged in e-negotiations reflect the negotiation traits and trends; Figure 1 shows an example from the beginning of a negotiation (Kersten et al, 2002). (Kersten and Zhang, 2003) used the

(Buyer) *Hi Joe, I'm Lisa and I represent Cypress Cycles in this negotiation. After extensive deliberation we have prepared an offer to purchase sprockets and gear assemblies. We think it is a fairly good offer and hope you find it acceptable.*
(Seller)*Hi Lisa, I am Joe, the representative of Itex Manufacturing and I am very delighted to get in touch with you. First of all, thank you very much for the possibility to negotiate with you and your company. Despite your really interesting offer, it is not possible for me and my company to accept it under all circumstances. For that reason I would like to make the following proposal to you. I am very interested in what you are thinking about, so I am looking forward to hearing from you. Bye, Joe.*

Figure 1: A sample of e-negotiation

history records of e-negotiations to study how the negotiation outcome depends on the intensity and distribution of offers exchanged during negotiation. However, such records and statistics might not be available in practice (esp. when, say for instance, the negotiation is not held via a negotiation support system). In such cases, the text used by the negotiators in their message-exchanges can prove to be useful. We examine this realm and hence work with the transcripts of the Inspire negotiations.

The *Inspire* text data (Kersten et al, 2002) is the largest text collection gathered through e-negotiations (held between people who learn to negotiate and may exchange short free-form messages). Negotiation between a buyer and a seller is successful if the virtual purchase has occurred within the designated time, and is unsuccessful otherwise. The system registers the outcome. We use the transcripts of 2557 negotiations, 1427 of them successful. We consider a transcript as a single example, with all messages concatenated chronologically, preserving the original punctuation and spelling. A successful negotiations is a positive example, an unsuccessful negotiation – a negative example. The *Inspire* data contain 27,055 word types

which constitute the initial feature set. That is, we apply feature selection to the data that contain 2557 examples and 27,055 features.

3 Feature-Selection Approaches

We want to compare two broad classes of feature selection methods and the feature subsets that these methods produce. As an evaluation criterion we use the results of the learning of classifiers on data represented using each of these feature subsets with respect to the outcome of negotiations.

We consider two *process-based* feature selection methods, negotiation-related and strategy-related, and two *corpus-based* methods, which represent the data with the most frequent unigrams and with *indicative* words. There is a major difference between the methods of the two classes. The former relies on expert knowledge about the domain from which the data originate. The latter requires feature scoring based purely on the statistical properties of the data. There is another difference: the extent of automation. *Process-based* approaches are inevitably semi-automatic, unlike the fully automatic *corpus-based* approaches that do not require integrating any expert knowledge.

3.1 Process-based Approaches

This type of feature selection is based on two different criteria. The *negotiation-related* feature selection identifies features specific to the process of negotiations. We can also build on the knowledge of influence-strategies that the negotiators employ. The features thus identified are called *strategy-related* feature selection.

Negotiation-related features (Shah et al, 2004) include words with specific negotiation-related meanings. Such words have been found to be unusually frequent compared to the typical word distribution in standard corpora. Selection of the negotiation-related features is based on the idea of identifying the elements of the communication model(Hargie and Dickson, 2002) of negotiations and works as follows:

- Consider the key elements of negotiations and identify these elements for the specific negotiations. Examples of such elements include: Environment (in the *Inspire* data – business), Goal (reaching an agreement), Topic (the purchase of good), Social roles within negotiations (buyers and sellers) and outside negotiations (students).

- Build the N-gram models from the data for N = 1, 2, 3.

- Identify semantic categories for the elements of negotiations; for example, the categories "hobbies" and "studies" can be identified for the social roles outside negotiations, the category "negotiation-specific" – for the goal, topic and environment.

- With respect to these categories, disambiguate each word – if necessary – using the most frequent bigrams and trigrams in which it appears.

- Build a semantic lexicon from the text data. Tag each word type[1] with one or more semantic category, using a lexical resource with semantic information (a machine-tractable form of (Summers, 2003)). In case of multiple candidate tags, select the one that corresponds to the most frequent sense of the word.

- Select the words tagged as negotiation-specific.

Strategy-related feature selection approach is based on the influence strategies most commonly used in negotiations. We present the general framework; see (Sokolova and Szpakowicz, 2005) for the details of the theoretical background and the implementation. To deliver the strategies, negotiators use appeal, logical necessity, and the indicators of intentions towards the subject of the negotiations and the negotiation process. In language, these strategic tools are exhibited in persuasion, substantiation, exchanges of offers, agreement and refusal (Brett, 2001); they reflect the reasoning, opinions and emotions of the participants. They are signalled by pronouns, negations, modal verbs, mental verbs, volition verbs and adjectives. Selection of the strategic features works as follows.

- Identify the influence strategies used in negotiations. *Direct* strategies are used when a negotiator directly influences the counterpart to

[1] A word type represents all occurrences of the same string in a text.

Negotiation-related features	
Word categories	Word types
nouns	offer, price, delivery...
action verbs	reduce, return, prepare...
volition verbs	agree, accept, refuse...
adjectives	recent, unacceptable...
mental verbs	think, know...

Table 1: Examples of negotiation-related features.

Strategy-related features	
Word categories	Word types
personal pronouns	I, we, you...
negations	no, none, nothing,...
modal verbs	can, will, should...
volition verbs	accept, promise, refuse...
adjectives	next, last, final, ...
mental verbs	think, understand, consider...

Table 2: Examples of strategy-related features.

make desirable concessions, *indirect* strategies – when attempts to influence the counterpart are not explicit.

- Represent influence strategies with the expression of persuasion, argumentation, substantiation, rejection and denial, and so on.

- Find a mapping between the word categories and the categories representing these strategies: negations are mapped to rejection and denial, modal verbs – to argumentation, mental verbs are associated with the intention towards the process of negotiations, and so on.

- Build the list of word categories including modals, volition verbs, negations, mental verbs, superlative adjectives. Finally, automatically extract from the data the words belonging to these categories.

Tables 1 and 2 give examples of negotiation-related and strategy-related features for the *Inspire* data[2].

3.2 Corpus-based Approaches

We evaluate the effectiveness of automatic corpus-based feature selection on two approaches. First, we use 200 most frequent unigrams counted in the e-negotiation corpora (one built from the data of successful negotiations, the other from the data of unsuccessful negotiations). These unigrams are chosen so that their frequencies are approximately the same in both successful and unsuccessful negotiations. With this set of features, we want to investigate if the features most frequently used in *both the negotiation classes* assist in binary classification. As opposed to most frequent words, *indicative* words are the unigrams whose frequency differs considerably in successful and unsuccessful negotiations. To identify these words we separate the data into two sets – successful and unsuccessful negotiations – and calculate the log-likelihood statistics LL for each word w (Rayson and Garside, 2000).

$$LL(w) = 2 * \left(\left(a * log\left(\frac{a*(a+b)}{c}\right)\right) + \left(b * log\left(\frac{b*(a+b)}{d}\right)\right)\right)$$

where a and c are the number of occurrences of w and the number of word tokens respectively, in the first corpus; b and d, in the second corpus. The higher the $LL(w)$, the larger the difference between frequencies of the word w in the two corpora.

3.3 The Datasets

For sets of features selected by each of the approaches described in subsections 3.1 and 3.2, we form bags of features from their unigrams. In each case, we build two datasets:

1. with the numerical attributes whose values are the numbers of occurrences of the word in negotiation; in this case we add one more attribute, whose value is the number of occurrences of other unigrams in the negotiation[3];

2. with the binary attributes showing whether the feature appears in the negotiation; there is no additional attribute.

4 Empirical Results

We have introduced several feature selection methods for e-negotiation. Now, we evaluate them using three learning paradigms. Paradigms with different

[2]The lists of negotiation-related features and strategic features intersect on seven features.

[3]To show that this attribute is relevant to the outcomes, we filter the attributes with Weka-based filters (Witten and Frank, 2000); this always selects the additional attribute as relevant.

learning biases give us an insight into the consistency of the results across them. We use C5.0, a version of C4.5 (Quinlan, 1993), a decision-tree learner that classifies entries by separating them into classes according to information gain of the attributes. Kernel methods, especially Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000), have been successfully used for text classification. They are also resistant to noise and work well on data with arbitrary distributions. We apply a linear kernel SVM. We also apply the probabilistic Naive Bayes classifier (NB) (Duda et al., 2000). NB was used with kernel density estimation and with the normal distribution estimation to model the numerical values (Witten and Frank, 2000). NB with kernel density estimation has shown better accuracy. We therefore report results only for NB with kernel density approximation.

We present tenfold cross-validation estimates of accuracy. To find out how the classifiers work on individual data classes, we use the standard text classification metrics: precision (P), recall (R) and equally-weighted F-measure. We have performed an exhaustive search on the adjustable parameters for every method. The classifiers were run on both sets of features: numerical, with the attribute values taking into account the frequency of occurrence of the selected set of features for each method; binary, with attribute values 0 for the absence and 1 for presence of the selected feature. Because of the identical performance of all classifiers on the sets of the most frequent and indicative features, we exclude the latter from the binary experiments.

Tables 3, 4 and 5 report the highest accuracy and corresponding P, R, F achieved by each classifier on every feature set and feature representation. For both numerical and binary representations, we *italicize* the highest accuracy for each classifier and put in ***bold*** the highest accuracy among them. The highest precision and recall are shown in **bold**. In our experiments, the baseline accuracy and precision are 55.8%, recall is 100%, and F-measure is 71.6% when we classify all negotiations as successful.

We do not present statistical significance because our results do not give enough material for a thorough *ANOVA* test for differences among groups; *ANOVA* would be the best method of exploring the difference in performance of combinations of the data features, their representation, and a classifier. *t-test*, used for a pair-wise comparison, is clearly not a suitable candidate. Additionally, Tables 4 and 5 show that the process-based features give the highest precision and recall for both numerical and binary representations. In the next section we explain how the process-based data representations affect the classification of positive and negative examples, that is, successful and unsuccessful negotiations.

5 The Informativeness of the Feature Sets

The features selected by the process-based approaches give higher classification accuracy than the features selected by the corpus-based approaches, but the two feature selection methods differ in what characteristics they extract from the data.

- The negotiation-related feature set is *specific* to negotiation; it captures the negotiators' main goal with respect to the negotiation issues, preferences and scope (width, depth, generality, specificity), and the numerical representation features reveal the intensity of the discussion of negotiation issues.

- The strategy-related feature set is *generic* in the sense that it does not relate specifically to negotiation issues; it rather captures the intentions to continue a negotiation, the influence on the partner, self-obligations and motivations, openness to feedback or the opposite, the boundaries within personal communication, and so on.

Negotiation-related and strategy-related features, although process-based, represent different aspects of the same process and therefore vary in their informative capacity. These differences allow learning of negotiation outcomes from various perspectives. Figures 2 and 3 report the true positive and true negative rates corresponding to the accuracies reported above. The results show that the negotiation-related features give higher accuracy in correct identification of positive examples and lead to the following explanation:
- the positive class either is homogenous or consists of a few well-represented subclasses;
- the negative class is divided into several small subclasses, and some of these subclasses are underrepresented.

Features	attr	NB	SVM	C5.0	attr	NB	SVM	C5.0
negotiation-related	num	*69.3*	71.7	***75.4***	bin	69.4	*74.0*	***74.8***
strategy-related	num	65.3	71.3	74.5	bin	*71.1*	72.7	73.7
most frequent	num	64.3	*73.4*	71.5	bin	64.2	71.5	73.3
indicative	num	64.2	72	74.4	bin	n/a	n/a	n/a

Table 3: Classification accuracy.

Features	# of attr	NB			SVM			C5.0		
		P	R	F	P	R	F	P	R	F
negotiation-related	124	72.3	72.5	72.5	72.5	75.8	74	73.3	**87.7**	79.9
strategy-related	100	74	58.3	56.7	**74.8**	73.2	74.0	72.5	87.6	79.3
most frequent	201	74.6	54.4	62.6	72.9	75.3	74.1	72.4	84.2	80.0
indicative	201	74.6	54.6	62.9	73.2	75.8	74.5	73.0	85.9	79

Table 4: Precision and recall; numerical representations.

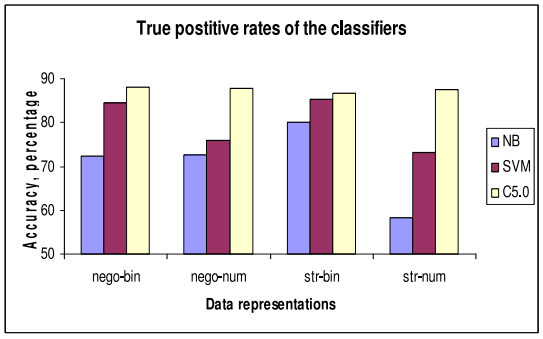

Figure 2: Classification of positive examples

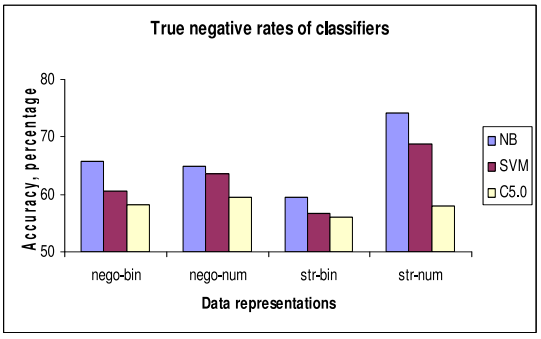

Figure 3: Classification of negative examples

This means that similarities among successful negotiations are easily revealed through the use of negotiation-related features and are strong enough to build a homogenous class, whereas for unsuccessful negotiations this assumption does not hold. The strategy-related features improve the classification accuracy by correctly identifying negative examples, especially when the binary representation is used. These features extract stronger similarities from the negative class than from the positive one. In the context of negotiations this suggests that discussing the topic of negotiation helps identify successful negotiations, while studying the implementation of influence-strategies helps identify unsuccessful negotiations.

We have shown that the two process-based approaches are complementary in the sense that they address different problems in learning from e-negotiation texts. It is natural to ask whether the benefits of both the sets of features can be exploited simultaneously. One possible direction of investigation would be to continue work with the features, either by constructing new ones, for example, building collocations of negotiation-related and strategy-related features, or suggesting an elaborate features selection method. Another opportunity to benefit from both sets of features comes from building an ensemble of classifiers, where the classifiers built by the same learner use different sets of features to classify the data and then combine their results. SVMs with the high accuracy and the most balanced performance on the data are the reasonable candidates.

6 Conclusion and future work

We have categorized, empirically compared and analyzed various feature selection and data represen-

Features	# of attr	NB			SVM			C5.0		
		P	R	F	P	R	F	P	R	F
negotiation-related	123	66.4	72.3	69.5	**73.1**	84.6	78.4	72.6	**88**	77.3
strategy-related	99	71.5	80.2	75.6	71.4	85.3	77.7	71.3	87.4	78.9
most frequent	200	74.6	54.6	63.1	72.9	75.3	74.2	72.3	84.2	77.8

Table 5: Precision and recall; binary representations.

tation methods for the text data collected during electronic negotiations. In particular, we compared two broad classes: the process-based and corpus-based feature selection methods. For each method from these two classes, we have studied their informativeness and representativeness with respect to the classification of the outcomes of e-negotiations. We have focused on the problem of identifying the learning settings that better assist the prediction of negotiation outcomes, where the settings include features, their representation and the learning paradigm. The classification of the negotiation outcomes was one of the evaluation measures.

We have shown empirically that the sets of features selected by the process-based approaches provide better classification of negotiation outcomes than the sets of features selected by the corpus-based approaches. We have confirmed this conclusion for NB, SVM and C5.0. Our analysis has shown that within the process-based feature selection approaches, the negotiation-related and strategy-related features complement each other on the classification of successful negotiations and unsuccessful negotiations. Thus, the features are good candidates for the future work on classification of the negotiation outcomes from texts.

The empirical results and their analysis should be helpful in work on knowledge-based electronic negotiation systems. We suggest the means of predicting the negotiation outcome and warning the negotiators when their language use may lead to the failure of negotiations.

Acknowledgments

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

References

Brett J. M. 2001. *Negotiating Globally*. Jossey-Bass.

Cristianini, N., J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press.

Duda, R., P. Hart, D. Stork 2000. *Pattern Classification*. Wiley, 2nd ed..

Hargie, O., D. Dickson. 2004. *Skilled Interpersonal Communication: Research, Theory and Practice*. Routledge, 4th ed.

Kersten G. E. and others. 2002-2006 *Electronic negotiations, media and transactions for socio-economic interactions*. interneg.org/enegotiation/.

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

Summers, D. (ed). 2003 *Longman Dictionary of Contemporary English*. Pearson Education: Longman. 4th ed.

Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Rayson, P., R. Garside. 2000. Comparing Corpora using Frequency Profi ling. Proc Workshop on Comparing Corpora (ACL 2000), 1–6.

Sebastiani, F. 2002 "Machine Learning in Automate Text Categoriazation". *ACM Computing Surveys*. 1-47.

Shah, M., M. Sokolova, S. Szpakowicz. 2004 "The Role of Domain-Specifi c Knowledge in Classifying the Language of E-negotiations". *Natural Language Processing (Proc of ICON'2004)*. 99–108.

Sokolova, M., S. Szpakowicz. 2005 "Classifi cation and Strategy Analysis in Electronic Negotiations". *Advances in Artificial Intelligence (Proc of AI'2005)*. 145–157.

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