Speech Act Based Classification of Email Messages in Croatian Language

Tin Franović, Jan Šnajder

University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{tin.franovic, jan.snajder}@fer.hr

Abstract

Speech acts provide an effective way of summarizing the intended purpose of an email message. In this paper we address the task of speech act classification of email messages in Croatian language. We frame the task as a multilabel text classification problem. We perform thorough evaluation using six machine learning algorithms on message-level, paragraph-level, and sentence-level features. Using message-level features, we achieved an overall best F1 score of over 94%.

Razvrščanje sporočil elektronske pošte v hrvaškem jeziku na podlagi govornih dejanj

Govorna dejanja predstavljajo učinkovit način za povzemanje namena sporočila elektronske pošte. V članku obravnavamo nalogo razvrščanja sporočil elektronske pošte v hrvaškem jeziku na podlagi govornih dejanj. Nalogo opredelimo kot problem razvrščanja besedila na podlagi oznak. Izvedemo poglobljeno evalvacijo z uporabo šestih algoritmov strojnega učenja in več nabori značilk na različnih ravneh – na ravni sporočila, na ravni odstavka ter na ravni stavka. Z uporabo značilk na ravni sporočila smo dosegli najboljši F1 izid preko 94%.

1. Introduction

The increase in popularity of email as means of business and personal communication is reflected in the amount of messages users are required to deal with on a daily basis. Recent surveys indicate that most people who use email for business purposes spend up to two hours a day reading, writing, and sorting email messages. This clearly indicates that there is a need for automated classification of email messages, which would drastically reduce the amount of time users spend on reading and sorting them. Classification of incoming email messages provides the user with information about the predicted importance or content of the messages before the user even reads them. This allows the user to focus on the messages considered important or interesting. Email classification has first become popular through *spam* filtering, which removes from the inbox the messages classified as unsolicited and places them in a special folder. Another solution is the filtering of messages classified as potentially important into a special folder called the *priority inbox*. Both techniques have successfully been implemented in widespread email clients.

The two aforementioned methods filter the messages based on their predicted importance. While importance-based filtering is convenient for most users, it is often difficult to predict what users will find important in a particular situation or context. The alternative to importance-based filtering is content-based classification, which labels each message based on its content, leaving it to the user to decide on the importance of the message.

This paper focuses on using speech acts expressed in email messages in the Croatian language for the purpose of content-based email classification. Speech acts are illocutionary acts that attempt to convey meaning from the speaker (or writer) to the listener (or reader) (Searle, 1965). In the context of email classification, speech acts provide an effective way of summarizing the intended purpose of the message. By labeling the email messages with speech acts which they contain, we enable the user to decide on which messages to focus first while reading. In this paper, we frame the speech act classification problem as a multilabel text classification problem and address it using supervised machine learning. We perform thorough evaluation experiments using six machine learning algorithms and three types of features extracted at three discourse levels (message, paragraph, and sentence level). We evaluate our speech act classifiers on a manually annotated collection of email messages in the Croatian language.

The rest of the paper is structured as follows. In the next section we give a brief overview of previous work on speech act classification. In Section 3 we describe our approach to speech act classification of email messages in Croatian. In Section 4 we evaluate the classifiers and discuss the results. Section 5 concludes the paper.

2. Related Work

The study of speech act classification (or *dialogue act classification*, as it is sometimes referred to) is one of the interesting challenges in natural language processing (NLP). From the NLP perspective, speech act classification is interesting especially for dialogue-based human-computer interaction. Successful dialogue systems are capable of understanding the speaker's intention and the message the speaker wishes to convey. Interpreting the speaker's intention is usually accomplished by analyzing and classifying speech acts. The *Clarity* project (Finke et al., 1998) is one of the first works in which speech acts are used in an attempt to understand dialogues. The focus of the project was to infer three levels of discourse structure in Spanish telephone conversations: speech acts, dialogue games, and discourse segments. The AutoTutor system (Marineau et al.,

2000) is an English computer tutor sensitive to speech acts from the previous dialogue turn, allowing the tutor to select the next action according to the speaker's intent. Keizer (2001) designed a conversational agent for the Dutch language that probabilistically interprets dialogue acts. Serafin et al. (2003) employ Latent Semantic Analysis (LSA) to classify speech acts from a corpus of tutoring dialogues in Spanish. Louwerse and Crossley (2006) use n-gram algorithms to classify speech acts from English dialogues on a location map reconstruction topic.

Relevant to the work presented in this paper is the use of speech acts for content-based email classification. Cohen et al. (2004) presented a system for classication of email messages in English based on supervised machine learning and a custom taxonomy of speech acts. In their subsequent work, Carvalho and Cohen (2006) exploit the linguistic aspects of the content-based classification problem by combining message preprocessing and n-gram feature extraction in order to improve the classification.

3. Speech Act Based Message Classification

3.1. Dataset annotation

There are several email datasets publicly available on the Internet, such as the *Enron* dataset (Klimt and Yang, 2004). However, none of these sets is in Croatian language. We therefore decided to first build a suitable dataset. An email dataset can essentially be obtained in two ways: by simulating a communication process (i.e., a business project communication), where different people take on different roles, as done by Cohen et al. (2004), or by finding a group of volunteers willing to provide their email messages sent over a period of time. We used the latter approach, mainly because volunteers were readily available and because the former method would take more time and resources. The total number of messages, collected from five sources, is 1337. Four sources contain personal emails provided by volunteers, while the fifth consists of messages exchanged during the course of a small student project.

For annotation, we used a set of 13 different speech acts, which can be divided into five groups according to Searle's classification (Searle, 1965):

- Assertives (AMEND, PREDICT, CONCLUDE);
- Directives (REQUEST, REMIND, SUGGEST);
- Expressives (APOLOGIZE, GREET, THANK);
- Commisives (COMMIT, REFUSE, WARN);
- Declarations (DELIVER).

The message annotation was split between two annotators, each annotating approximately one half of the dataset. The annotators were asked to annotate in each email portions of text that contain a speech act. The size of these portions may vary from a few words to larger portions spanning over several sentences. As a general rule, one speech act annotation could not span over multiple paragraphs. The total number of messages is 1337, the number of paragraphs is 4468 paragraphs and number of words 76,760. The number of annotated speech acts in the dataset is 4498, and for different speech acts the number of annotations is between 14 for the REFUSE speech act and 1069 for the GREET speech act. On average, a speech act annotation contains 17.06 words, with CONCLUDE being the longest on average (32.1 words), while GREET being the shortest (5.99 words per annotation).

Table 1: κ-statistic for all speech acts

Speech act	κ	Speech act	κ
AMEND	0.714	REFUSE	0.000
APOLOGIZE	0.856	REMIND	0.747
COMMIT	0.851	REQUEST	0.589
CONCLUDE	0.005	SUGGEST	0.544
DELIVER	0.792	THANK	0.949
GREET	0.779	WARN	0.174
PREDICT	0.267		

Table 2: Classifier performance on speech acts (% F1)

	NB	k-NN	SVM	DS	AB	RDR
DELIVER	69.70	83.72	88.16	85.71	87.50	**88.51**
AMEND	**79.31**	71.43	77.97	72.29	74.63	77.27
COMMIT	62.45	67.44	78.61	79.37	81.97	**83.75**
REMIND	60.87	63.64	75.00	76.92	**94.74**	76.92
SUGGEST	67.06	70.27	**76.84**	76.27	75.12	71.50
REQUEST	69.69	75.44	**78.76**	70.57	75.23	74.46

The two annotators double-annotated 15% of the dataset, on which we evaluated the inter-annotator agreement. The κ statistic (Carletta, 1996), computed separately for each speech act, is shown in Table 1. On some speech acts (REFUSE, CONCLUDE, WARN) the agreement was considerably low, thus we decided to exclude these speech act from further consideration. After removing the infrequent speech acts and speech acts with low inter-annotator agreement, we ended up with six speech acts: DELIVER, AMEND, COMMIT, REMIND, SUGGEST, and REQUEST. The removed speech acts are: APOLOGIZE, CONCLUDE, GREET, PREDICT, REFUSE, THANK, and WARN.

3.2. Message preprocessing

Message preprocessing consisted of stop-word removal, stemming, and the extraction of training examples. We created a separate training set for every speech act. Using the information provided by each annotation (original message, start and end point of annotation), we extracted text segments corresponding to the sentence, paragraph, and message levels. At the message level, we use the whole original message text. At the paragraph and sentence levels, we extract the text segments that enclose the start and end points of the annotation. If an annotation spans over multiple sentences, all of the sentences are included. Negative examples for every speech act are sampled from the set of text segments not annotated with the corresponding speech act. The number of negative examples was chosen to be approximately the same as the number of positive examples.

In order to reduce the dimensionality of the input space and eliminate the morphological variation, we applied a

Table 3: Classifier performance on discourse levels (% F1)

	Message	Paragraph	Sentence
DELIVER	86.59	83.64	**88.51**
AMEND	**79.31**	77.27	72.38
COMMIT	**83.75**	81.97	78.93
REMIND	**94.74**	76.92	69.57
SUGGEST	71.88	**76.84**	69.74
REQUEST	70.09	**78.76**	72.19
Overall	94.74	83.64	78.93

Table 4: Overall classifier performance (% F1)

	Message	Paragraph	Sentence
NB	**79.31**	69.70	72.38
k-NN	72.73	75.44	**83.72**
SVM	83.87	81.55	**88.16**
DS	78.65	79.37	**85.71**
AB	**94.74**	83.54	87.50
RDR	86.59	83.64	**88.51**

simple stemming procedure: we removed the the suffix of each word after the last vowel (including the vowel itself) if the length of the suffix is less than half the length of the word. Stemming reduced the number of terms from 15,100 to 11,856. Apart from stemming, we optionally employ stop-word filtering. Stop-words are common function words that, in the context of content-based text classification, are usually filtered out because they carry little semantic information. We used a list of 2024 Croatian stop words.

3.3. Training classifiers

For the classification experiment we use Rapid Miner, an open-source data mining environment that simplifies the training process and provides a variety of classifiers to choose from. We experiment with six different models: SVMs (Support Vector Machines), naive Bayes (NB), k-NN (k-Nearest Neighbors), Decision Stump (DS), AdaBoost (with Decision Stump as the weaker learner), and RDR (Ripple Down Rule). For all models, apart from RDR, we experiment with two term weighting schemes: TF (Term Frequency) and TF-IDF (Term Frequency – Inverted Document Frequency). For RDR, we use binary weights in order to obtain interpretable rules, which are based on the presence or absence of a term in a message. We train a separate classifier for every speech act, term weighting scheme, and discourse level. Because we are considering six speech acts, three term weighting schemes (one for RDR and the other two for the other models), three discourse levels, and a total of six different classifier types, the total number of models trained is 198. Additionally, we have trained all models using feature sets with reduced dimensionality obtained by removing the stop-words.

For training and validation we used 70% of the dataset, while the remaining 30% we used as a held-out test set. The training process includes the optimization of model parameters (except for NB and DS, which have no model parameters), which we accomplished using grid-search and cross-validation. For every parameter combination, a model is trained and evaluated using 10-fold cross-validation and the optimal parameters are chosen based on the F1 score averaged over ten folds. The optimal model is then re-trained on the whole training set and evaluated on the held-out set.

4. Evaluation

4.1. Classifier performance

Table 2 shows the performance of the six classifiers on the six different speech acts in terms of the F1 score. Here we show the performance of the best-performing models regardless on the discourse level or features used. The SVM and RDR classifiers consistently outperform other considered classifiers, with F1 scores reaching over 88%. SVMs not only performed well, but also had the lowest difference between the best and the worst performance, ranging from 75% (for the REMIND speech act) to 88.16% (for the DELIVER speech act). AdaBoost also showed a consistently good performance, and was the best performing classifier for the REMIND speech act. DS showed surprisingly good results, considering the simplicity of the model.

It can also be seen that most of the classifiers perform best on the DELIVER speech act. On the other hand, the REMIND speech act proved to be the most difficult to classify, which may be attributed to the fact that this speech act had by far the lowest number of training examples.

4.2. Discourse level

Table 3 shows the classifier performance on the three different discourse levels. We again show the performance of best-performing classifiers, regardless of features used.

The results exhibit no particular global regularities, such as that better performance may be obtained on sentences rather than on the complete messages, as might have been expected. However, the results may help us understand what are the levels on which particular speech acts are usually expressed. For instance, a reminder to someone is rarely expressed with a single sentence, thus it would be expected to see that for this particular speech act a classifier performs better on the message or paragraph level. On the other hand, deliveries are usually expressed in a small number of words, which is why classification at the sentence level showed to perform the best.

Overall, classification at the message level has shown to perform best for most speech acts, followed by the paragraph level. This could be attributed to the fact that all classifiers have a very high recall, and more surrounding text is needed to filter out the false positives.

4.3. Features

Table 5 shows the results obtained by choosing the best-performing classifier for each pair of speech act and feature type. In general, stop-word removal seems not to influence the classification performance. In the case when there is no stop-word removal, the performance of all three feature types was comparable in that there is no consistent pattern where one feature type outperforms the others, with only the binary feature under-performing for the REQUEST

Table 5: Classifier performance with respect to feature types (% F1)

	With stop-words			Without stop-words		
	Binary	TF	TF-IDF	Binary	TF	TF-IDF
DELIVER	**88.51**	87.50	88.00	**88.51**	88.16	87.96
AMEND	70.07	77.19	**79.31**	77.27	75.86	**77.19**
COMMIT	**83.75**	79.37	81.63	78.82	79.76	**81.97**
REMIND	76.92	76.92	**77.78**	75.00	**94.74**	77.78
SUGGEST	71.50	**76.84**	76.27	68.40	73.08	**73.68**
REQUEST	61.90	**78.76**	78.10	74.46	**78.08**	77.53

speech act. The differences between the F1 scores using different feature types were usually confined within 3%, which shows that the problem at hand is generally robust with respect to the term weighting schemes used.

4.4. Overall performance

The best performance of each classifier for a particular discourse level is presented in Table 4. Most classifiers show their best performance on the sentence level, which is in contradiction with the observation that for most speech acts the best classification is achieved on the message level. This, however, can be explained by taking into account that these results are highly influenced by the very high performance of all classifiers on the DELIVER speech act. The overall performance of the classifiers is relatively high compared to reported results for the English language: our F1 scores range from 79.31% to 94.74%, whereas Cohen et al. (2004) report F1 scores from 44% to 85%.

5. Conclusion

Speech acts provide an effective way of summarizing the intended purpose of email messages. We addressed the task of speech act classification of email classification in Croatian language. We framed this task as a multilabel text classification problem and performed thorough evaluation using six machine learning algorithms and three types of features (message-level, paragraph-level, and sentence-level features). We have shown that the discourse level and feature type do not significantly influence the performance. However, we were able to demonstrate that certain speech acts are more accurately classified at a particular discourse level. Using message-level features, we achieved an overall best F1 score of over 94%. The obtained F1 scores are notably higher than those reported in previous work.

An issue that we have not addressed in this paper is the practicaly usability of speech act classification for importance-based email classification; we leave this investigation for future work. Also for future work, we intend to further explore the relationship between the discourse levels and the speech act. Another possible direction of research would be to employ information extraction methods to augment each speech act with additional information such as named entities, temporal expressions, etc.

6. Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia under Grant 036-1300646-1986. We thank the anonymous reviewers for their comments.

7. References

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, pages 249–254.

V. R. Carvalho and W. W. Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, pages 35–41.

W. W. Cohen, V. R. Carvalho, and T. M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of EMNLP 2004*, pages 309–316.

M. Finke, M. Lapata, A. Lavie, L. Levin, L. Mayfield Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. 1998. Clarity: Inferring discourse structure from speech. In *In Proc. of Workshop on Applying Machine Learning to Discourse Processing*.

S. Keizer. 2001. A Bayesian approach to dialogue act classification. In *BI-DIALOG 2001: Proc. of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pages 210–218.

B. Klimt and Y. Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, pages 217–226.

M. M. Louwerse and S. A. Crossley. 2006. Dialog act classification using n-gram algorithms. In *FLAIRS Conference*, pages 758–763.

J. Marineau, P. Wiemer-Hastings, D. Harter, B. Olde, P. Chipman, A. Karnavat, V. Pomeroy, A. Graesser, and the Tutoring Research Group. 2000. Classification of speech acts in tutorial dialog. In *Proc. of the workshop on modeling human teaching tactics and strategies, Intelligent Tutoring Systems 2000*, pages 65–71.

J. R. Searle. 1965. What is a speech act? *The Philosophy of Language, Oxford University Press*, pages 44–46.

R. Serafin, B. Di Eugenio, and M. Glass. 2003. Latent semantic analysis for dialogue act classification. In *Proceedings of HLT-NAACL 2003–short papers*, volume 2 of *NAACL-Short '03*, pages 94–96, Stroudsburg, PA, USA. Association for Computational Linguistics.