Data-Driven Identification of Dialogue Acts in Chat Messages

Dietmar Schabus, **Brigitte Krenn**, and **Friedrich Neubarth**
Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
firstname.lastname@ofai.at

Abstract

We present an approach to classify chat messages into dialogue acts, focusing on questions and directives (“to-dos”). Our multi-lingual system uses word lexica, a specialized tokenizer and rule-based shallow syntactic analysis to compute relevant features, and then trains statistical models (support vector machines, random forests, etc.) for dialogue act prediction. The classification scores we achieve are very satisfactory on question detection and promising on to-do detection, on English and German data collections.

1 Introduction

Online chat systems are a form of text-based communication that has been available since the early days of the Internet and that has become widespread in a variety of uses. In recent years, chat systems as a tool for business-internal communication seem to be an especially active market and such systems are sometimes replacing e-mail as the primary means of written communication within organizations. Uthus and Aha (2013) provide a survey over the active research field of automatic chat analysis. In processing chat messages, we face the challenge that traditional Natural Language Processing (NLP) techniques often do not work well, as with other forms of *microtext* (Ellen, 2011). This is due to typical characteristics such as message brevity, incorrect and/or non-standard spelling and grammar, fragment sentences, lack of or non-standard usage of punctuation, as well as influences from spoken and face-to-face communication for expressing sentiment or emphasis (e.g., emoticons, emojis, and character repetitions).

In this paper, we address the problem of classifying chat messages according to *dialogue acts* (Stolcke et al., 2000), i.e., the problem of assigning each chat message to one of a few categories that reflect the role of the message within the (multi-party) dialogue. For our purposes, the term dialogue act is roughly equivalent to the older term *speech act* (Searle, 1969). Typical dialogue acts of interest are various question types and directives, and in particular their realization in chat-based company communication.

Tagging chat messages with dialogue act labels can be useful as an intermediary step for other tasks such as thread disentanglement (Shen et al., 2006; Uthus and Aha, 2013), for facilitating information retrieval and extraction on chat data (e.g., search result filtering based on dialogue acts), and for enabling the chat platform itself to offer “smart” features based on dialogue act tags, for example.

We present an approach on dialogue act tagging of chat messages based on data-driven classification techniques, including random forests and support vector machines, which are trained on a collection of business chat logs. We show that our approach reaches high classification accuracy in an experimental evaluation. The system combines language specific and language independent components. In the current paper we present results for English and German.

The remainder of this paper is organized as follows. Section 2 presents related work, Section 3 describes the data corpus we work with and the dialogue acts we want to identify. In Section 4 we describe our method in detail, and evaluate its performance in Section 5. Finally, Section 6 draws conclusions and points out some ideas for future work.

2 Related Work

Wu et al. (2005) define 15 dialogue acts, including statement, yes-no-question and wh-question, to classify chat messages, using transformation-based learning with expert-provided rule templates. They report to achieve F_1 scores of 0.70 for yes-no-questions and 0.53 for wh-questions. Using the same 15 categories on a different corpus, Forsyth and Martell (2007) compare a neural network to to a naive Bayes classifier. As features they use several message distances and occurrence counts of specific keywords, as well as presence of certain words as the first token of the message. For the neural network they report F_1 scores of 0.75 for yes-no-questions and 0.74 for wh-questions.

To detect questions in discussion threads on Yahoo! Answers, Wang and Chua (2010) use sequential pattern mining and syntactic shallow pattern mining (parse trees, to which a simplification procedure is applied) as features for a one-class support vector machine. They report an F_1 score of 0.91.

Carpenter and Fujioka (2011) define 43 dialog act categories and use long string matching and several rules (starts with, ends with, contains) to classify IRC chat messages. They report 90% accuracy, but state that this is partly due to the constrained context of the messages in their corpus.

Both Dent and Paul (2011) and Li et al. (2011) attempt to detect questions in Twitter messages, using different rule sets. They achieve F_1 scores of 0.71 and 0.92, respectively. The latter paper also evaluates an approach to detect interrogatives based on support vector machines, however it did not result in an improvement of detection accuracy. Zhang et al. (2011) also work with Twitter messages, but categorize them using five dialogue act categories, including statements and questions. By training a support vector machine on unigram, bigram and trigram features, they achieve an F_1 score of 0.64 both for the question category and as an overall average.

Kim et al. (2010) detect 12 dialogue acts (including open questions, yes-no-questions and requests) in one-on-one chats using conditional random fields on bag-of-words features and additionally exploiting structural and inter-utterance dependencies. They have later expanded their work to 14 dialogue acts on multi-party chats (Kim et al., 2012), where they report F_1 scores of 0.42, 0.75 and 0.87 for requests, wh-questions and yes-no-question, respectively.

O'Shea et al. (2013) attempt to distinguish questions from non-questions, using decision trees trained on 22 part-of-speech-like categories of function words as features, with sentences represented as category vectors. They report classification accuracies of 99% on their "straightforward question vs. non-question without preamble" data set and 79% on their "simulated clauses" data set.

Category	Ab.	Examples
Wh-quest.	wh	@ron so whats the state of dev now
Y-N-quest.	yn	or can I just specify one of them
Echo quest.	ec	a username can contain slashes?
Non-quest.	nq	just read your post :P
Directive	td	nice, please send them to me
Non-dir.	nd	lol no problem. ^^

Table 1: Sample utterances for the categories considered in question and directive classification.

It should be noted that all mentioned contributions deal with English data only, whereas we work on both English and German data and have already generalized many aspects of our system to work with multiple languages. Furthermore, all above papers employ either a rule-based approach or a machine learning approach on very simple features. We extract relevant syntactic features using a small rule set and then employ machine learning techniques. As we show in Section 5, our syntactic features are crucial for the classification results we achieve. In addition to the detection of questions, which many others have also investigated, we also detect directives, which are much less commonly considered.

3 Data Sets

Focusing on the detection of two groups of dialogue acts, *questions* and *directives*, we have assembled collections of sample sentences for classifier training and testing both in English and German. In question detection we attempt to classify any given message as either a *wh-question* (based on an interrogative word), a *yes-no-question* or a *non-question* (e.g. a declarative statement). Additionally, we use the label *echo question* for questions that do not exhibit clear interrogative grammatical structure, such as declarative statements ending in a question mark and fragments with a question mark.

In directive detection we intend to distinguish directives ("to-dos") from non-directive messages. Directives are often phrased as imperatives, but note that it is also possible that a given utterance is both a directive and a question, e.g., "Can you write the report, please?". Table 1 gives examples from our data for all categories under investigation.

Each of the four subcorpora comprises 1500 hand-labeled utterances, which were taken from

Class:	nq	yn	ec	wh	total	td	nd	total
English	618	379	261	242	1500	501	999	1500
German	819	233	204	244	1500	679	821	1500

Table 2: Class frequencies of the question corpora for English and German (left) and of the directive corpora for English and German (right).

various sources, including real English business chat messages, provided to us by our project partner,[1] real German chat messages from the "Dortmunder Chat-Korpus" (Beißwenger, 2013) and sentences/utterances taken from out-of-copyright novels in English and German.[2] By using this mixture we attempt to cover both typical chat message style as well as more grammatically rigid and more elaborate language from novels. Many chat messages in our collection are very short, and even in the longer ones complex sentence structures are very rare. For each question class, we have removed the question marks from 50% of the utterances that originally had one, such that the classifier cannot rely on the presence or absence of question marks alone.

The class frequencies of the four subcorpora are given in Table 2. Note that we have separate disjoint data collections for the question detection task and the directive detection task, i.e., we have 6000 labeled utterances altogether. Ideally, the numbers for the two languages would be more symmetrical, but as we do not focus on a comparison between them, we consider this no serious problem. The agreement between two labelers was higher than 98% for both question subcorpora and higher than 84% for both to-do subcorpora. This difference is due to the fact that the definition of to-dos is by far not as clear-cut as that of questions. Sometimes it can only be decided on a semantic or pragmatic level, assuming a certain context, whether or not a given chat message should be labeled as a to-do or not. We therefore also expect our automatic classifiers to perform better on questions than on to-dos.

4 Method

We have developed a software pipeline for dialogue act detection in chat messages with support for multiple languages. Most parts of the pipeline are language independent, the few language specific ones are currently available for English and German. Both in the training phase and later during detection, messages are first split into utterances and tokens using a custom tokenizer we have developed for chat messages. The tokenizer uses English and German lexica with more than 400,000 and 2,000,000 full form entries, respectively. Looking up a given lexeme in the lexicon yields all possible readings wih the repective part-of-speech (POS) tags and morpho-syntactic features (e.g., "bears" yields a plural noun reading and a third person singular verb reading).

Given that standard NLP tools such as POS-taggers and parsers would not work well on the short and fragmented utterances typically found in chat, especially without adequate training data, we refrain from applying such techniques. Instead we operate on ambiguous morpho-syntactic information as retrieved from the lexicon on which we perform a shallow rule-based analysis: Starting from the beginning of the message, we skip all tokens that are greetings, interjections, conjunctions, adpositions or non-words like URLs, emoticons etc. The first token that is not to be skipped is labeled p_1 (intuitively, the first syntactically relevant word in the message). Starting from p_1, and depending on the (possible) morpho-syntactic features of the token at p_1, a small rule set continues to skip tokens that may belong to the syntactic phrase headed by p_1. After that, the next token is labeled p_2, for example:

$$\textit{haha well ok but } \underbrace{\textit{which}}_{p_1} \textit{ of these things } \underbrace{\textit{are}}_{p_2} \textit{ true?}$$

When this heuristc procedure works well, p_1 will point to the subject of the clause and p_2 to the finite verb in a declarative statement, and vice-versa in a yes-no-question. In a wh-question, p_1 will point to the interrogative pronoun and p_2 to the finite verb, etc. Position p_2 or even both p_1 and p_2 may be undefined, for example when the message consists only of a single interjection. With this simple procedure for shallow syntactic analysis, we are able to capture the most relevant structural properties for detecting questions and directives, even in short and incomplete sentences.

Given the (ambiguous) POS tags and other morpho-syntactic features for each token, as well as the two positions p_1 and p_2, we define a high-dimensional binary feature vector, which contains, amongst others: The POS tags, lemmata and morpho-syntactic features at p_1 and p_2, all of

[1] `http://grape.io`

[2] `https://www.gutenberg.org`

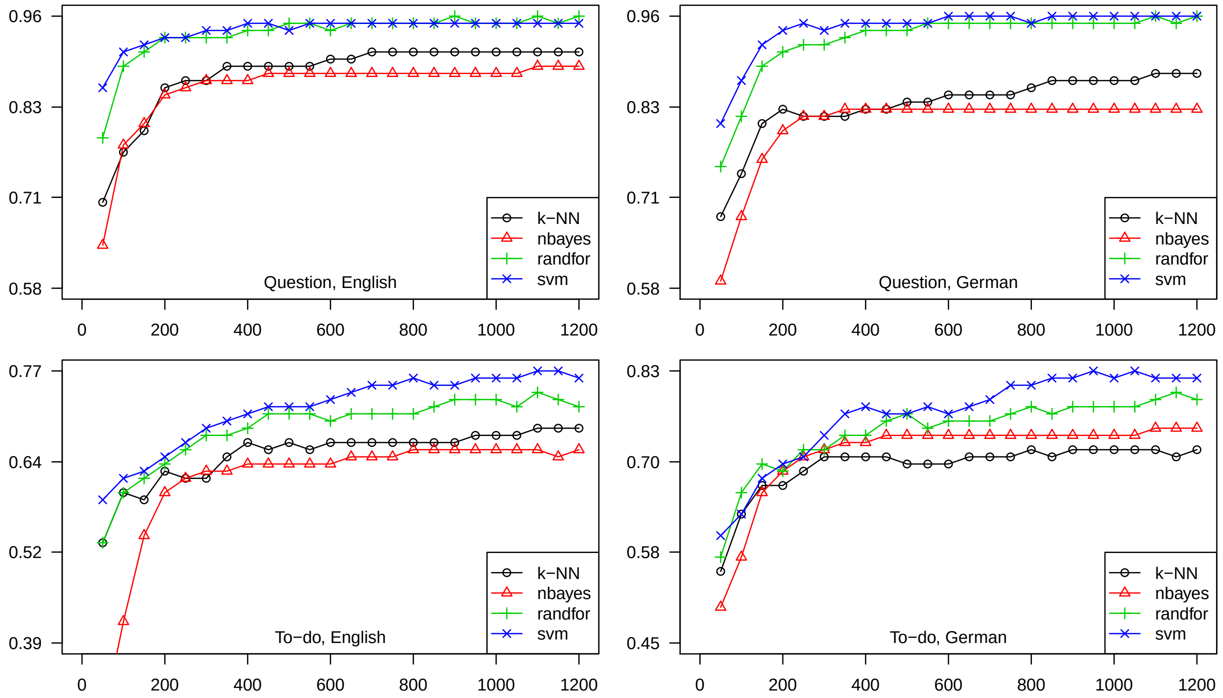

Figure 1: Evaluation results for the four data sets. The horizontal axis shows to the number of training samples. The vertical axis shows the average F_1 score across five cross validation folds. All four plots have identical vertical scaling.

these features appearing anywhere in the utterance, and the presence of some indicative phrases (e.g., “please”, “can you”, “you should”). It should be noted that POS, lemma and morpho-syntactic features are ambiguous for many tokens, as described above. All features are encoded as binary variables in the feature vector indicating the presence or absence of a certain feature (such as “noun at p_1” and “plural at p_2”).

The features and data described above are used to train a classification model such as a support vector machine or random forest. The following section investigates the performance of various methods. After training, new input messages can be classified by first detecting the language of the input, applying utterance splitting, tokenization, rule-based syntactic analysis and feature extraction as described, and finally by using the model to predict the dialogue act.

5 Evaluation

To evaluate the method described in Section 4 on the data described in Section 3, we have carried out a series of experiments. For each of the four data sets of 1500 utterances, a five-fold cross validation setup was employed, yielding 1200 training utterances and 300 test utterances per fold. Within each fold, we initially used only 50 utterances to train a model and gradually increased this number to the full 1200, always evaluating the model on the same 300 test utterances. The whole procedure was repeated using the following four modeling approaches: k-nearest neighbors (k-NN; $k = 5$), naive Bayes (nbayes), random forest (randfor; 100 trees) and support vector machine (svm; linear kernel) from the scikit-learn library (Pedregosa et al., 2011). The results are shown in Figure 1, where each data point is an average F_1 score across the five folds, and in the case of the multi-class problem of question detection also across the four classes (macro averaging).

We observe that overall better results are achieved in question detection than in to-do detection, as expected. Interestingly, the results of the two best methods (svm and randfor) begin to level off already around 500 training utterances for question detection, but they continue to rise for to-do detection, suggesting that in the latter case additional training data could further improve the results.

The results for the best method (svm) using all the 1200 utterances for each fold are shown in Table 3, which lists precision, recall and F_1 score for each of the classes. For question detection, all val-

	English			German		
Cl.	Prec.	Rec.	F_1	Prec.	Rec.	F_1
nq	0.961	0.953	0.957	0.977	0.978	0.977
ec	0.969	0.957	0.963	0.977	0.960	0.968
wh	0.911	0.933	0.922	0.975	0.983	0.979
yn	0.966	0.972	0.969	0.919	0.916	0.917
td	0.783	0.750	0.765	0.804	0.834	0.818

Table 3: Precision, Recall and F_1 score values per category resulting from 5-fold cross validation using a linear kernel support vector machine.

	English			German		
Cl.	Prec.	Rec.	F_1	Prec.	Rec.	F_1
nq	0.855	0.878	0.865	0.942	0.950	0.946
ec	0.827	0.716	0.765	0.805	0.794	0.796
wh	0.764	0.837	0.796	0.838	0.888	0.862
yn	0.709	0.695	0.701	0.709	0.647	0.673
td	0.719	0.599	0.653	0.711	0.699	0.702

Table 4: Precision, Recall and F_1 score values per category resulting from 5-fold cross validation using a linear kernel support vector machine, when the features based on p_1 and p_2 are not used.

ues are above 0.95 except for English wh-questions and for German yes-no-questions, where they are still above 0.91. To-do detection is less reliable, here all values are greater than 0.75.

Typical errors made by the system include: free relative clauses that are mistaken for a wh-question (“What strikes me is that ...”), statements with dropped subject pronoun that are mistaken for a to-do (“love it!”, “just read your post”), yes-no-questions with dropped auxilary verb that are not recognized correctly (“you on your way?”), and to-dos that our system misses because they are expressed indirectly (“john, the build system needs an update”) or phrased in a way that is too complex for our simple approach (“I would like to note that you still need to finish the presentation”).

Interestingly, if we remove the features based on the p_1 and p_2 positions, we observe a substantial drop of the classification results, as shown in Table 4. For example, recall drops from 0.972 to 0.695 for English yes-no-questions and from 0.750 to 0.599 for English to-dos; similar for German. This large difference indicates that our shallow syntactic analysis is crucial for the good classification results we achieve.

6 Conclusion

We have presented a data-driven approach for classifying chat messages into dialogue acts, with a focus on (several types of) questions and directives, in English and German. We use (ambiguous) POS and other morpho-syntactic information in combination with a rule-based shallow syntactic analysis as features for several learning algorithms, with support vector machines achieving the best results in our experiments. Our F_1 scores for question detection seem better than those in related work, although a fair comparison would require a standardized evaluation corpus. For a problem that has not received a lot of attention, our scores in to-do detection are also promising, with some room for improvement. The shallow syntactic analysis plays a key role in our system; in future work we plan to make this component also data-driven rather than rule-based. Furthermore, we would like to additionally consider the conversational context of each message for improving the detection of to-dos.

Acknowledgments

This research was funded in part by the Austrian Research Promotion Agency (FFG) in the “FFG Basisprogramm” under project number 850380. The authors would like to thank UberGrape GmbH (http://grape.io) for the fruitful collaboration and for providing interesting real-world data for us to work with.

References

Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, 41(1):161–164, April.

Tamitha Carpenter and Emi Fujioka. 2011. The role and identification of dialog acts in online chat. In *Proceedings of the Workshop on Analyzing Microtext at the 25th AAAI Conference on Artificial Intelligence*, pages 2–7, San Francisco, CA, USA, August.

Kyle Dent and Sharoda Paul. 2011. Through the Twitter glass: Detecting questions in micro-text. In *Proceedings of the Workshop on Analyzing Microtext at the 25th AAAI Conference on Artificial Intelligence*, pages 8–13, San Francisco, CA, USA, August.

Jeffrey Ellen. 2011. All about microtext – a working definition and a survey of current microtext research within artificial intelligence and natural language processing. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence (ICAART)*, pages 329–336, Rome, Italy, January.

Eric N. Forsyth and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC)*, pages 19–26, Irvine, CA, USA, September.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 862–871, Cambridge, MA, USA, October.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2012. Classifying dialogue acts in multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 463–472, Bali, Indonesia, November.

Baichuan Li, Xiance Si, Michael R. Lyu, Irwin King, and Edward Y. Chang. 2011. Question identification on Twitter. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2477–2480, Glasgow, UK, October.

James D. O'Shea, Zuhair A. Bandar, and Keeley A. Crockett. 2013. Optimizing features for dialogue act classification. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 474–479, Manchester, UK, October.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John R. Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread detection in dynamic text message streams. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–42, Seattle, WA, USA, August.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

David C. Uthus and David W. Aha. 2013. Multiparticipant chat analysis: A survey. *Artificial Intelligence*, 199–200:106–121.

Kai Wang and Tat-Seng Chua. 2010. Exploiting salient patterns for question detection and question retrieval in community-based question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1155–1163, Beijing, China, August.

Tianhao Wu, Faisal Khan, Todd Fisher, Lori Shuler, and William Pottenger, 2005. *Foundations of Data Mining and knowledge Discovery*, volume 6, chapter Posting Act Tagging Using Transformation-Based Learning, pages 319–331. Springer, Berlin/Heidelberg, August.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2011. What are Tweeters doing: Recognizing speech acts in Twitter. In *Proceedings of the Workshop on Analyzing Microtext at the 25th AAAI Conference on Artificial Intelligence*, pages 86–91, San Francisco, CA, USA, August.