Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pages 161–162.
Trento, Italy, 30 May – 1 June 2007. Edited by Ron Artstein and Laure Vieu.

Complex Taxonomy Dialogue Act Recognition with a Bayesian Classifier

Mark Fishel
Dept of Computer Science
University of Tartu
Tartu, Estonia
fishel@ut.ee

1 Introduction

This paper describes the experiments of performing dialogue act (DA) recognition with a complex DA taxonomy using a modified Bayes classifier.

The main application of DA recognition is in building dialogue systems: classifying the utterance and determining the intention of the speaker can help in responding appropriately and planning the dialogue. However, in this work the target application is human communication research: with tagged DAs it is easier to search for utterances of a required type in a dialogue corpus, to describe the dialogues with a general model of dialogue moves, etc.

The DA taxonomy, used in the current work, was designed for the Estonian Dialogue Corpus (EDiC) (Hennoste et al., 2003). This means two additional difficulties for DA recognition. Firstly, DA taxonomies used for human communication research are as a rule much more detailed than in case of dialogue systems (e.g., comparing DCIEM (Wright, 1998) and CallHome Spanish (Ries et al., 2000) taxonomies); therefore, more DAs have to be distinguished, with several of them having unclear meaning boundaries. Secondly, Estonian is an agglutinative language with 14 cases, a complex system of grouping and splitting compound nouns, heterogeneous word order and several other features that make natural language processing harder.

2 Experiments

2.1 Experiment Setup

In order to determine the optimal set of input features additive feature selection was applied. All of the tests were performed using 10-fold cross-validation.

In this work we only tried simple features, not involving morphological analysis, part-of-speech tagging, etc. The used ones included DA tag bi- and trigrams, keywords and the total number of words in the utterance. Keyword features included the 1st word, first 2 words and first, middle and last words as a single dependency. We also tried stemming the words and alternatively leaving only the first 4 characters of the word.

The learning model used in this work is the Bayes classifier. Its original training/testing algorithm supports only a fixed number of input features. This makes it harder to include information with variable size, such as the set of the utterance words. In order to overcome this limitation, we slightly modified the algorithm by calculating the geometrical average of the conditional probabilities of the DA tag, given each utterance word. With this approach the probabilities remain comparable despite the variable length of the utterances.

The corpus used for training and testing is described in greater detail in (Gerassimenko et al., 2004), updated information can be found online[1]. The version used in the experiments contains 822 dialogues (a total of 32860 utterances) of mixed content (telephone conversations in an information service, at a travelling agency, shop conversations, etc).

2.2 Results

After the feature selection process converged, the following features were included into the selection:

[1] http://math.ut.ee/~koit/Dialoog/EDiC

DA tag trigram probabilities, the geometrical mean of the word-tag conditional probabilities and the number of words in the utterance. Stemming was not performed in the final preprocessing.

The resulting cross-validation precision over the whole set of dialogues was 62.8% with the resulting feature set. In general the most typical DA tag to be confused with was the most frequent one. In addition, some tags were frequently confused with each other.

In addition to the objective precision estimation provided by cross-validation, we also wanted to have a direct comparison of the resulting DA tagger with the human taggers. For that we applied the tagger to both human tagged parts, used for calculating the human agreement. The resulting precisions for the two parts are 80.5% and 78.6%.

3 Discussion

It is interesting to note that the resulting selection of representation features included only simple text-based features. Although the task of DA recognition belongs to computational pragmatics in natural language processing, in this case it gets solved on the level of pure text, which is even lower than the morphology level.

Future work includes several possibilities. In particular, several output errors of the trained classifier seem obvious to solve to a human tagger. For instance, several utterances containing wh-words are misclassified as something other than wh-questions. There are at least two possibilities to treat that kind of problems. Firstly, a set of rules can be composed by professional linguists to target each output problem individually. This approach has the advantage of guaranteed improvement in the required spot; on the other hand, manually composing the rules can result in overlooking some global influences on the remaining utterance cases, which can cause decreased performance in general. Another way to address the output errors would be to add more descriptive features to the input.

4 Conclusions

We have described a set of experiments, aimed at applying a Bayes classifier to dialogue act recognition. The targeted taxonomy is a complex one, including a large number of DA tags.

Additive feature selection was performed to find the optimal set of input features, representing each utterance. The tested features included n-gram probabilities and keyword-based features; the latter were tested both with and without stemming.

The resulting precision of the trained model, measured with 10-fold cross-validation is 62.8%, which is significantly higher than previously achieved ones. The selected features included DA tag trigram probabilities, number of words probability and the geometrical mean of the word-tag conditional probabilities of all the utterance words.

The model was compared to the agreement of human taggers in the targeted taxonomy – this was done by applying it to the same test corpus that was used in calculating the agreement. The two resulting precisions are 80.5% and 78.6%, which is very much near the human agreement (83.95%).

There is much room for further development of the classifier. This includes adding more specific features to the model's input, manually composed output post-processing rules, etc.

References

Olga Gerassimenko, Tiit Hennoste, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo, and Evely Vutt. 2004. Annotated dialogue corpus as a language resource: An experience of building the estonian dialogue corpus. In *Proceedings of the 1st Baltic Conference "Human Language Technologies. The Baltic Perspective"*, pages 150–155, Latvia, Riga.

Tiit Hennoste, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo, and Evely Vutt. 2003. Developing a typology of dialogue acts: Tagging estonian dialogue corpus. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue, DiaBruck 2003*, pages 181–182, Saarbrücken, Germany.

Klaus Ries, Lori Levin, Liza Valle, Alon Lavie, and Alex Waibel. 2000. Shallow discourse genre annotation in callhome spanish. In *Proceecings of the International Conference on Language Ressources and Evaluation (LREC-2000)*, Athens, Greece.

H. Wright. 1998. Automatic utterance type detection using suprasegmental features. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, volume 4, page 1403, Sydney, Australia.