DIALOGUE ACT CLASSIFICATION USING LANGUAGE MODELS

Norbert Reithinger and Martin Klesen *

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
E-Mail: reithinger@dfki.de

ABSTRACT

Pragmatically important information as e.g. dialogue acts that describe the illocution of an utterance depend in traditional processing approaches on error prone syntactic/semantic processing. We present a statistically based method for dialogue act classification that has word strings as input. An experimental evaluation shows that this method can be successfully used to determine dialogue acts. The overall recognition rate in the experiments is in the range of 65%–67% for German test data, and 74% for an experiment with English dialogues.

1. INTRODUCTION

For spoken or typed language processing systems, it is often important to extract just the core intention of a sentence. The classes are sometimes described as dialogue moves [3] or, as in the VERBMOBIL system, as dialogue acts [1]. In the latter system which translates spoken time scheduling dialogues of two humans [2] these dialogue acts classify sentences e.g. as ACCEPT, REJECT, SUGGEST. They are used e.g. in a shallow translation component within VERBMOBIL.

Currently, we have 43 acts defined, including approximately 10 which describe deviations from the task. The acts are organised in a hierarchy where the nodes describe domain independent illocutions and the leaves contain domain specific information.

A common knowledge base in VERBMOBIL is a corpus of spoken and transliterated scheduling dialogues. More than 800 of them (German and English) have been annotated with dialogue related information and serve as the empirical foundation for different methods within dialogue processing, e.g. the statistical prediction of dialogue acts [7]. Each utterance of this corpus has been hand-annotated with a dialogue act. An evaluation of the reliability of this annotation is presented in [7]. It is demonstrated that the dialogues are consistently labeled and can be used as a reliable basis for processing.

Traditional linguistic approaches compute these classes from syntactically and semantically rich feature structures which require extensive, hand-coded linguistic and world knowledge sources. However, a system built in such a traditional way has to cope with the problem of cumulative error. Each processing module has only a certain success probability. In a simplified view cumulative error is then the product of all modules which contributed to processing. In the VERBMOBIL case, the speech input is processed and transformed by 3 modules until information about the dialogue act reaches the dialogue module which needs these acts as basic input. Therefore, even if each module has a 90% success probability, the cumulative error for the input of the dialogue module is as high as 28%, assuming statistical independency.

In this paper we present an approach that needs no expensive and error-prone deep processing. We first demonstrate the algorithms that are based on dialogue act specific language models and evaluate this method on three test sets of dialogues.

2. THE ROBUST DIALOGUE ACT CLASSIFICATOR

The task to be solved is to decide which dialogue act D describes the illocution of a certain string of words W most probable, i.e. to solve the classification problem

$$D = \operatorname*{argmax}_{D'} P(D'|W)$$

which can be as usual reformulated as

$$D = \operatorname*{argmax}_{D'} P(W|D')\ P(D')$$

(see also [6]). The a-priori probabilities $P(W|D')$ can be approximated from the annotated corpus. It is split according to the annotated dialogue acts, resulting in 43 partitions. These can be used for the approximation of the probabilities of strings by computing relative frequencies. We evaluated numerous methods for approximation like linear and rational interpolation, backing–off or using the multivariant Poisson distribution as proposed by Garner et.al. [3]. However, the best results are provided by using a linear interpolation of uni- and bigrams.

*This work was funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01IV101K/1. The responsibility for the contents of this study lies with the authors.

Additionally, instead of using just $P(D)$ as statistical dialogue model, we can exploit the knowledge about the previous dialogue history H by using $P(D'|H)$ from the dialogue act prediction component [7]. Experiments show that correct classifications are up to 3% better if we include this additional parameter. Therefore, the classificator we use is

$$D = \underset{D'}{\operatorname{argmax}}\ P(W|D')\ P(D'|H)$$

3. EVALUATION

We evaluated the method in numerous experiments. Main evaluation criteria are precision, defined as

$$\texttt{precision} = \frac{\texttt{correct classifications}}{\texttt{actual classifications}}$$

and recall, defined as

$$\texttt{recall} = \frac{\texttt{correct classifications}}{\texttt{possible classifications}}$$

The performance of the approach is shown in more detail for three experiments. For the first two experiments, the training data consist of 350 German dialogues and the test data comprises 87 German dialogues.

For the first experiment, we tried to recognise all 43 dialogue acts. Table 1 shows recall (*rec*) and precision (*prec*) for those dialogue acts that occur more than 5 times in the test data. The table is sorted according to recall.

As can be seen, dialogue acts that are realized with some fixed phrases like GREET or BYE, are recognised with a very high recall and precision. But also those dialogue acts that are important for the progression of the negotiation like most SUGGESTs, ACCEPTs, and REQUESTs have at least a recall and precision rate of 60% to 80%. The most frequent one, namely SUGGEST_SUPPORT_DATE — 610 utterances in the test corpus — has a recall of 83.93%, with 64.81% precision. Overall, from the 2912 utterances in the test corpus, 1898 dialogue acts are recognised correctly, i.e. 65.18%

Especially those dialogue acts which describe deviations from the task like DIGRESS_SCENARIO are usually not recognised very well. Reasons are that they either comprise very inhomogeneous linguistic material or are infrequent in the training corpus. The latter reason is also true for some of the acts with a LOCATION/DURATION propositional content.

There are also some dialogue acts like SUGGEST_EXCLUDE_DATE which are not clearly separated from other acts, e.g. from REJECT_DATE.

The dialogue acts are organised in a hierarchy with additional 18 acts describing primarily intentions at a domain independent level [4]. Exploiting this hierarchy it is possible to map the domain dependent acts in the annotated corpus to the 18 acts. Then, we can train language models

dialogue act	*rec*	*prec*
THANK_INIT	96.77	90.91
GREET	95.12	89.31
INTRODUCE_NAME	92.54	89.86
BYE	87.34	87.34
SUGGEST_SUPPORT_DATE	80.98	66.94
REQUEST_COMMENT_DATE	79.28	78.57
REJECT_DATE	71.49	71.19
ACCEPT_DATE	71.0	64.69
INIT_DATE	69.9	67.29
REQUEST_SUGGEST_DATE	68.42	61.18
GIVE_REASON	61.81	64.96
FEEDBACK_ACKNOWLEDGEMENT	56.14	44.14
SUGGEST_EXCLUDE_DATE	46.94	41.44
DELIBERATE_EXPLICITE	46.34	62.3
SUGGEST_SUPPORT_LOCATION	44.44	42.11
REQUEST_SUGGEST_DURATION	40.0	100.0
FEEDBACK_RESERVATION	36.96	50.0
CLARIFY_STATE	35.71	43.75
CONFIRM	33.33	37.5
CLARIFY_QUERY	32.73	60.0
DIGRESS_SCENARIO	26.47	38.3
ACCEPT_LOCATION	26.32	50.0
MOTIVATE_APPOINTMENT	25.0	16.67
SUGGEST_SUPPORT_DURATION	22.22	57.14
GARBAGE	18.67	42.42
CLARIFY_ANSWER	18.6	53.33
DELIBERATE_IMPLICITE	18.18	43.48

Table 1: Results for all dialogue acts that occur in the corpus at least 5 times (German)

for the more abstract classes. Table 2 shows the results for the same training and test dialogues as used in the first experiment. The column *pos* gives the number of possible utterances in this class, *act* the number of actual classifications for this class, and *cor* the number of correct classifications.

Again, the dialogue acts realised primarily by fixed phrases are recognised best. While REQUEST and SUGGEST are pretty stable, the precision for ACCEPT gets better about 2%. Overall, from the 2912 utterances in the test corpus, 1951 dialogue acts are recognised correctly, i.e. 67.18% This is approx. 2% better than the result from the experiment comprising 43 dialogue acts.

This table shows also that – as to be expected – the dialogue acts are not equally distributed. SUGGESTions are by far the most frequent acts in the corpus (approx. 25%), followed by ACCEPTs and REJECTs. However, material that doesn't contribute to the negotiation per se also has a big share in the test corpus, namely more than 15%.

A final experiment was made to show the usability of this approach for languages different from German. Form our annotated English corpus, we used 143 dialogues as training material, and 20 as test corpus.

Table 3 shows the results for this experiment. Although the test corpus is relatively small it can be seen that comparable results as with the German dialogues are achieved. Acts realized as phrases are recognised very good, and the negotiation-relevant ones are also pretty good in their respective recall and precision. Overall, from the 328 utterances in the test corpus, 245 dialogue acts are recognised

dialogue act	*rec*	*prec*	*pos*	*act*	*cor*
THANK	96.88	83.78	32	37	31
GREET	94.31	89.92	123	129	116
INTRODUCE	90.0	88.73	70	71	63
BYE	86.71	87.82	158	156	137
REQUEST_COMMENT	79.46	74.79	112	119	89
SUGGEST	79.03	67.28	744	874	588
REJECT	71.43	71.43	238	238	170
ACCEPT	71.19	67.11	427	453	304
REQUEST_SUGGEST	67.07	62.5	82	88	55
INIT	65.14	65.74	109	108	71
GIVE_REASON	61.81	62.68	144	142	89
FEEDBACK	50.31	45.76	161	177	81
DELIBERATE	37.23	61.45	137	83	51
CONFIRM	37.04	43.48	27	23	10
CLARIFY	32.14	50.4	196	125	63
DIGRESS	27.54	41.3	69	46	19
MOTIVATE	25.0	14.29	8	14	2
GARBAGE	16.0	41.38	75	29	12

Table 2: Results for the abstract classes (German)

dialogue act	*rec*	*prec*	*pos*	*act*	*cor*
THANK	100.0	100.0	3	3	3
DELIBERATE	100.0	100.0	2	2	2
BYE	100.0	94.12	32	34	32
REQUEST_SUGGEST	100.0	72.73	8	11	8
GREET	100.0	80.0	4	5	4
REQUEST_COMMENT	87.5	93.33	16	15	14
SUGGEST	81.6	76.69	125	133	102
ACCEPT	77.78	77.78	45	45	35
REJECT	74.36	70.73	39	41	29
INIT	66.67	44.44	6	9	4
DIGRESS	50.0	75.0	6	4	3
CLARIFY	40.0	33.33	5	6	2
GIVE_REASON	27.78	45.45	18	11	5
MOTIVATE	14.29	50.0	7	2	1
GARBAGE	14.29	100.0	7	1	1
FEEDBACK	0.0	0.0	2	4	0
CONFIRM	0.0	0.0	3	2	0
INTRODUCE	0.0	0.0	0	0	0

Table 3: Results for the abstract classes (English)

correctly, i.e. 74.7%. This is 7% better than for German. Even if the relative small test corpus does not allow for final conclusions, one reason for this higher recognition might be the fixed word order in English as compared to the relatively free phrase order of German.

The three experiments shown above are made using the transliterations of spoken dialogues. For a first evaluation of the method working on word lattices, we used 77 German lattices computed by one of VERBMOBIL's speech recognisers. The turns consist of utterances that stick to the topic of time scheduling. We extracted the best word chain, using the trigram language model of VERBMOBIL. The training data is, as above, 350 dialogues from the transliterated and annotated corpus.

The evaluation could not be directly against the transliterated and annotated corpus. One reason is that speech input in VERBMOBIL is usually not limited to one sentence and turns have to be segmented using prosodic information about sentence boundaries. These segments are very often different from the transliteration and additionally contain word recognition errors. We computed the dialogue acts for the segments, using the 18 domain independent acts, and evaluated the dialogue acts according to the dialogue act coding manual. A manual evaluation shows that for about 70% of the segments, the computed dialogue act was acceptable according to the annotation manual.

4. CONCLUSION

We have demonstrated the use of dialogue act dependent language models for the recognition of dialogue acts. Using a Bayesian approach, and including the dialogue history in the dialogue model, we implemented a classifier that uses word strings as inputs and computes the most probable dialogue act. For three experiments we showed that the dialogue acts that are realised mostly with fixed phrases have recall and precision percentages between 85% and 100%. Recall and precision for the negotiation relevant dialogue acts is in the 60%–65% range. As can be expected, dialogue acts that describe deviations from the topic can only be recognised poorly. The evaluation with English dialogues demonstrates that the method is working reliable for this language, too.

REFERENCES

[1] Jan Alexandersson, Norbert Reithinger, Elisabeth Maier: Insights into the Dialogue Processing of VERBMOBIL. In Proc. ANLP 97. Washington, DC, 1997.

[2] Thomas Bub, Johannes Schwinn: VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System. In Proc. ICSLP 1996, p. 2371-2374. Philadelphia, PA., 1996.

[3] Phil Garner, Sue Browning, Roger Moore, Martin Russell: A Theory of Word Frequencies and its Application to Dialogue Move Recognition. In Proc. ICSLP 1996, p. 1880-1883. Philadelphia, PA., 1996.

[4] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, J. Joachim Quantz: Dialogue Acts in VERBMOBIL. VERBMOBIL Report No. 65. Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.

[5] Elisabeth Maier: Evaluating Transcribed Speech Data Annotated with Discourse Information. VERBMOBIL Report No. 193. DFKI GmbH, Saarbrücken 1997.

[6] Marion Mast, Heinrich Niemann, Elmar Nöth, Ernst Günter Schukat-Talamazzini: Automatic Classification of Dialog Acts with Semantic Classification Trees and Polygrams. In: IJCAI-95, Workshop on Machine Learning. Montreal, 1995

[7] Norbert Reithinger, Ralf Engel, Michael Kipp, Martin Klesen: Predicting Dialogue Acts for a Speech-To-Speech Translation System. In Proc. ICSLP 1996, p. 650-663. Philadelphia, PA., 1996