

Leave One Out Experiments for Statistical Dialogue Act Recognition

Norbert Reithinger

DFKI GmbH

June 2000

Norbert Reithinger
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
Tel.: (0681) 302 - 5346
Fax: (0681) 302 - 5341

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01IV101K/1 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den AutorInnen.

ISSN 0949-6084

Leave One Out Experiments for Statistical Dialogue Act Recognition

Norbert Reithinger

June 15, 2000

Abstract

We describe a corpus based statistical approach to dialogue act recognition used in a speech translation system. We present the annotated corpus used for training and test purposes, the statistical method, and the results of leave one out experiments.

With this memo, we finish our work on statistical dialogue act recognition in Verbmobil.

1 Introduction

Computing intentions for sentences or utterances is important for a wide variety of tasks in natural language processing systems: from providing the building blocks for plan based descriptions of texts or dialogues to being an instrumental part of well-known speech translation systems like Janus or Verbmobil. In the 90s of the last millennium, computing these intentions, henceforth called dialogue acts, shifted from manually built knowledge-based approaches to classification approaches trained on annotated corpora. The numbers of approaches that were presented in the previous years show that these methods are now in the mainstream and are widely applied [Mast *et al.*, 1995, Reithinger and Klesen, 1997, Kipp, 1998, Samuel *et al.*, 1998, Tanaka and Yokoo, 1999, Choi *et al.*, 1999].

This paper summarizes the statistical dialogue act recognition method as implemented in Verbmobil [Wahlster, 2000], a speaker independent speech-to-speech translation system for German, English, and Japanese. Its domains are scheduling, travel planning, and hotel reservation. We first present some details about the corpus. The size and structure of the available training and test

material is important and influences the performance of the algorithms. We then recapitulate our recognition algorithm, and show its performance with leave one out experiments, using two different types of input.

2 Some Facts About the Verbmobil Corpus

Over the last years, more than 50 CD-ROMs of spoken dialogues were collected within the Verbmobil project. The core data are the audio files and the transliterations, i.e. transcripts with additional information about noises, overlaps, and similar features. This corpus is used for training of e.g. the speech recognizers. The transliterations are also the basis for various types of annotation, e.g. tree-banks or dialogue acts.

The dialogue act annotation was done at our site, including the definition of the annotation scheme [Alexandersson *et al.*, 1998]. Over the life-span of the project, we revised the scheme three times while the domain was extended from the initial scheduling domain to the current ones. Also, there was a significant learning curve that is mirrored in the final annotation manual. We will not discuss the process of annotation here — for an overview of guidelines collected in the MATE project see [Klein, 1999]— but we want to stress the importance of revisions, to keep an annotated corpus alive and usable.

<i>language</i>	# CD-ROMs	# dial.	# <i>dial. acts</i>	<i>mean</i>	<i>min</i>	<i>max</i>
German	13	738	37954	51.24	6	208
English	6	375	22682	59.93	7	347
Japanese	2	402	15574	39.52	16	83
Sum	21	1505	76210	50.41	6	347

Table 1: Annotated CD-ROMs

Table 1 shows the annotated CD-ROMs (# *CD-ROMs*), the number of annotated dialogues (# *dial.*) and dialogue acts (# *dial.acts*), and the mean, minimal, and maximal length of dialogues, measured in dialogue acts.

Figure 1 shows the dialogue length distribution for the three languages. On the x-axis, the length of dialogues in numbers of dialogue acts is shown, on the y-axis the number of dialogues with a certain length is drawn. As can be seen, German dialogues have a maximum around the length of 30 acts and few of the dialogues are longer than 100 acts. The absolute maximum for English is also around 30 acts, however the corpus contains quite a few dialogues that are really long. The Japanese corpus is much more homogeneous centered around

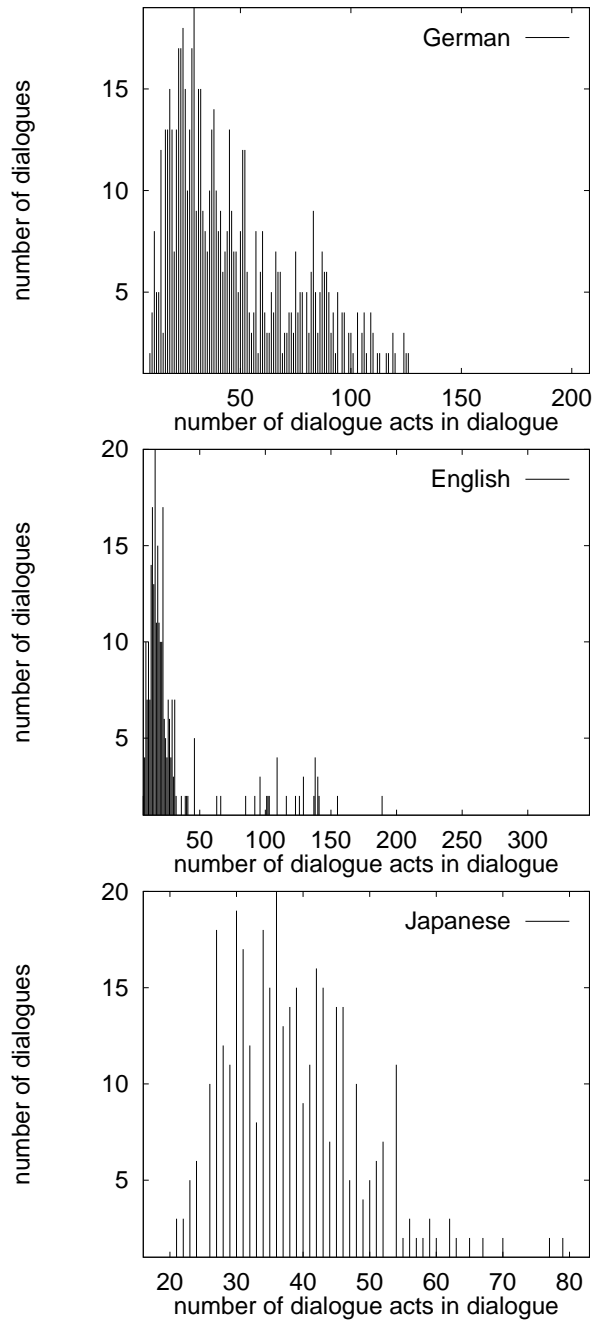


Figure 1: Distribution of dialogue length in the annotated corpus

the mean dialogue length. We will show below, how the length of the dialogues

influences the recognition results.

Our dialogue act scheme used for the annotation defines 35 dialogue acts, structured in a hierarchy [Alexandersson *et al.*, 1998]. However, for processing we use only 19 acts. Utterances annotated with the other 16 acts are mapped on the 19 used acts exploiting the hierarchy. The reasons for this approach are twofold: firstly, 10 dialogue acts combined cover less than 1% of the annotated utterances, and contain acts like `DEVIATE_SCENARIO`. They neither carry propositional content central to the negotiation dialogues nor do they control the dialogue like the act `DEFER` which occurs also not very frequently but has an important function in the progress of a dialogue. The other 6 acts are frequently confused with closely related acts, e.g. `FEEDBACK_POSITIVE` with `ACCEPT`. The main distinction in the annotation manual between the two acts is that the latter contains a propositional content, whereas the first one doesn't. We developed tools to create and analyse confusion matrices amongst others to detect such cases.

Table 2 contains the details about the dialogue acts: the total number and percentage ($\#(\%)$), the mean utterance length in words (*mul*), the minimal (*min*) and maximal (*max*) length in words, and the mean word length in characters (*mwl*). The Japanese data were annotated using the second revision of our dialogue act annotation scheme. This did not include the dialogue acts `CLOSE`, `COMMIT`, `DEFER`, `INFORM_FEATURE`, and `REQUEST_COMMIT`. The Roman transcription is based on syntactical and morphological criteria. Phrases (“bunsetsu”) are separated by spaces and are considered as words for further processing.

The annotated corpus is not only used to train a dialogue act recognizer, as described in the next section. It also is used for the training of the segmentation algorithms within the prosody module of *Verbmobil* [Nöth *et al.*, 1999]. The segmentation defined through the dialogue acts annotated is used for the modeling of one type of boundaries, namely sentence boundaries. In the running *Verbmobil* system, possible boundaries and their probabilities are annotated at each word in the word lattices which are generated by the speech recognizers. Modules which consume lattices consider these values when splitting a lattice into utterances.

3 The Recognition Method

In the past years, mainly three methods for classification of dialogue acts were proposed and applied to larger corpora

- statistical classifiers, using language models (see e.g. [Mast *et al.*, 1995,

dialogue act	All Data					German				
	# (%)	mul	min	max	mw	# (%)	mul	min	max	mw
ACCEPT	18472 (24)	3.99	1	55	4.154	8643 (23)	4.39	1	52	4.401
BYE	2722 (4)	2.97	1	23	5.391	1621 (4)	2.84	1	23	5.838
CLOSE	469 (1)	7.37	1	33	4.330	399 (1)	7.49	2	33	4.451
COMMIT	407 (1)	9.29	1	37	4.333	249 (1)	8.67	1	37	4.623
DEFER	217 (0)	11.55	3	36	4.219	143 (0)	10.55	3	24	4.571
GIVE_REASON	2746 (4)	11.10	1	49	4.302	1438 (4)	9.68	1	49	4.748
GREET	1880 (2)	3.34	1	15	4.539	1407 (4)	3.63	1	15	4.456
INFORM	13066 (17)	7.79	1	54	4.227	6002 (16)	7.52	1	52	4.763
INFORM_FEATURE	2674 (4)	11.13	1	58	4.346	843 (2)	10.44	1	41	5.195
INIT	2980 (4)	13.54	1	64	4.548	1702 (4)	12.43	1	37	5.139
INTRODUCE	1925 (3)	5.19	1	22	4.764	709 (2)	5.78	1	22	4.166
POLITENESS_FORMULA	1768 (2)	5.52	1	27	4.533	342 (1)	5.79	1	19	4.418
REJECT	5062 (7)	8.28	1	55	4.309	3084 (8)	7.89	1	55	4.604
REQUEST	2762 (4)	7.79	1	43	4.213	870 (2)	7.01	1	25	4.742
REQUEST_COMMENT	2360 (3)	5.84	1	77	4.041	1177 (3)	5.62	1	33	4.143
REQUEST_COMMIT	93 (0)	8.80	1	21	4.577	62 (0)	8.15	3	21	4.881
REQUEST_SUGGEST	2061 (3)	8.73	1	33	4.196	946 (2)	7.84	1	26	4.577
SUGGEST	13862 (18)	11.11	1	50	4.543	7876 (21)	10.63	1	50	5.019
THANK	684 (1)	3.23	1	16	4.909	441 (1)	3.11	1	10	4.702

dialogue act	English					Japanese				
	# (%)	mul	min	max	mw	# (%)	mul	min	max	mw
ACCEPT	6557 (29)	2.83	1	55	3.890	3272 (21)	5.25	1	41	3.893
BYE	625 (3)	3.59	1	19	3.270	476 (3)	2.62	1	11	7.561
CLOSE	70 (0)	6.71	1	15	3.562	n.a. (n.a.)	n.a.	n.a.	n.a.	n.a.
COMMIT	158 (1)	10.27	1	31	3.948	n.a. (n.a.)	n.a.	n.a.	n.a.	n.a.
DEFER	74 (0)	13.49	4	36	3.688	n.a. (n.a.)	n.a.	n.a.	n.a.	n.a.
GIVE_REASON	491 (2)	11.47	3	49	3.794	817 (5)	13.39	3	39	3.997
GREET	276 (1)	1.89	1	7	4.122	197 (1)	3.27	1	10	5.533
INFORM	4446 (20)	7.16	1	54	3.716	2618 (17)	9.50	1	44	3.906
INFORM_FEATURE	1831 (8)	11.45	1	58	3.989	n.a. (n.a.)	n.a.	n.a.	n.a.	n.a.
INIT	701 (3)	11.90	2	42	3.910	577 (4)	18.79	2	64	3.886
INTRODUCE	35 (0)	4.11	1	7	3.722	1181 (8)	4.87	2	21	5.215
POLITENESS_FORMULA	246 (1)	4.19	1	27	3.435	1180 (8)	5.72	1	19	4.734
REJECT	1311 (6)	8.58	1	33	3.737	667 (4)	9.54	1	37	4.190
REQUEST	986 (4)	6.86	1	35	3.905	906 (6)	9.56	2	43	4.081
REQUEST_COMMENT	551 (2)	5.15	1	30	3.650	632 (4)	6.87	2	77	4.140
REQUEST_COMMIT	31 (0)	10.10	1	17	4.086	n.a. (n.a.)	n.a.	n.a.	n.a.	n.a.
REQUEST_SUGGEST	452 (2)	8.71	1	33	3.784	663 (4)	10.02	3	31	4.015
SUGGEST	3735 (16)	10.32	1	42	4.011	2251 (14)	14.09	3	47	3.934
THANK	106 (0)	2.49	1	16	4.284	137 (1)	4.19	2	11	5.692

Table 2: Distribution of dialogue acts and length information

Reithinger and Klesen, 1997, Tanaka and Yokoo, 1999, Choi *et al.*, 1999])

- neural networks [Kipp, 1998]
- transformation based learning [Samuel *et al.*, 1998]

After evaluating all three approaches (see the comparison section below), we use the first one. It decides which dialogue act D describes the illocution of a

certain string of words W with statistical models for each D , with the condition of a given W . It selects the model that is most probable, i.e.

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

which can as usual be reformulated with Bayes' rule as

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

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 a prediction component [Reithinger *et al.*, 1996]. Experiments show that correct classifications are up to 3% better if we include this additional parameter. In the experiments presented below, recall and precision raise between 1% and 2% when using the dialogue history. Therefore, the classifier we use is

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

The a-priori probabilities $P(W | D')$ and $P(D' | H)$ can be approximated from the annotated corpus. All utterances for one dialogue act are collected and the relative word and transition frequencies are used for the approximation of $P(W | D')$. The dialogue act transition probability $P(D' | H)$ is also approximated by the relative transition frequencies of the acts in the dialogues.

In the literature, various methods are proposed to approximate the distributions. We implemented a flexible classifier workbench in LISP and augmented it with evaluation and visualization tools to analyse the results. Using this workbench, we tested interpolation algorithms [Klesen, 1997] like linear and rational interpolation, backing-off or using the multi variant Poisson distribution [Garner *et al.*, 1996]. However, the best results were provided when using a linear n -gram interpolation using equally distributed interpolation weights. The optimization of the weights using the EM-algorithm on an evaluation set usually reduces perplexity for this set but yields also an over-adaptation to it. Recognition rates in the test described below are better without this optimization.

4 Recognition Rates Using Words

Most approaches in dialogue act recognition partition their corpus in fixed sets of training and test dialogues (e.g. [Mast *et al.*, 1995, Reithinger and Klesen, 1997, Kipp, 1998, Samuel *et al.*, 1998, Choi *et al.*, 1999]). However, the selection of the

test/training partition might influence the results of the evaluation. Therefore, we made extensive leave one out experiments. In these experiments, we trained for each of the 1505 dialogues language models where only the test dialogue was held back from the training data.

As evaluation criteria we use recall, precision and the κ value which is used to compute the agreement between two coders [Carletta, 1996]

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the probability for the agreement between two coders and $P(E)$ the probability for agreement by chance. Trained human coders can reach κ values greater than 0.8. In the plots we multiply κ with 100. However keep in mind that κ values have a range from -1 to 1 and do not express a percent value.

The first experiment is concerned with the question how the n in the n -gram interpolation influences the results. We ran 9 leave one out test series, starting with just unigrams (relative frequency of words), interpolation of uni- and bigrams, and so on up to $n = 9$. Figure 2 shows for the three languages German, English, and Japanese, how recall, precision and κ change as longer n -grams are used for the interpolation. The values from the 1505 tests for each n -gram are summed up to overall numbers.

For German, English and Japanese, recall and κ – which depends on recall – have a peak for bigrams. Precision for German and English has a peak at 5-grams, for Japanese at 4-grams. In Japanese precision is always higher than recall, while it's the other way round in German and English.

Table 3 shows detailed recall and precision values for the dialogue acts, using bigram models and collecting the results of a whole leave one out experiment using all 1505 dialogues. Again, the numbers from all runs are summed up and finally recall and precision for the whole experiment is computed. As to be expected, the acts expressed by highly standardized phrases like GREET are recognized with a high accuracy. Problematic are acts like CLOSE where only few training samples are available. Acts that are confused frequently with each other like GIVE_REASON with REJECT can be identified using confusion matrices. If the classifier is used in a running system, modules consuming the data from such a classifier must be aware of this uncertainty.

To see how the length of our dialogues measured in dialogue acts is related to recall, we set up another leave one out experiment, using bigram models. In figure 3 we sorted the dialogues with respect to their length on the x-axis and draw the cumulated recall values, i.e. overall recall of the dialogues up to the length as indicated on the x-axis on the y-axis.

German			English		
<i>dialogue act</i>	<i>recall</i>	<i>prec.</i>	<i>dialogue act</i>	<i>recall</i>	<i>prec.</i>
GREET	96.30	94.36	THANK	96.23	97.14
THANK	95.92	91.96	BYE	95.20	89.21
INTRODUCE	95.20	95.88	GREET	94.93	94.58
BYE	93.83	91.35	ACCEPT	89.73	82.20
ACCEPT	83.82	76.21	INTRODUCE	77.14	93.10
INIT	76.32	67.83	POLITENESS_FORMULA	74.80	82.14
REJECT	73.91	68.46	SUGGEST	71.49	64.38
SUGGEST	73.07	69.94	REJECT	67.35	62.05
COMMIT	71.08	65.07	COMMIT	66.46	61.40
REQUEST_SUGGEST	68.82	65.17	INFORM_FEATURE	63.79	58.34
REQUEST_COMMENT	62.79	67.99	INIT	60.77	58.28
POLITENESS_FORMULA	55.85	57.70	REQUEST_SUGGEST	56.64	55.41
INFORM_FEATURE	51.84	46.00	REQUEST_COMMENT	51.54	61.21
CLOSE	51.13	36.43	INFORM	49.69	60.01
GIVE_REASON	49.37	55.79	DEFER	36.49	42.86
REQUEST_COMMIT	43.55	67.50	CLOSE	35.71	47.17
INFORM	40.05	52.76	REQUEST_COMMIT	32.26	66.67
DEFER	39.86	38.51	REQUEST	30.63	45.69
REQUEST	23.71	42.04	GIVE_REASON	25.25	35.23
<i>overall</i>	69.45	65.84	<i>overall</i>	68.46	66.16

Japanese		
<i>dialogue act</i>	<i>recall</i>	<i>prec.</i>
INTRODUCE	99.15	96.46
BYE	97.27	93.72
THANK	97.08	95.68
INIT	93.41	81.17
GREET	86.29	63.20
ACCEPT	85.70	76.51
POLITENESS_FORMULA	73.81	86.41
SUGGEST	72.99	64.38
REQUEST_SUGGEST	67.12	67.84
REJECT	66.42	56.01
REQUEST_COMMENT	66.14	61.93
GIVE_REASON	47.74	50.98
INFORM	44.88	57.82
REQUEST	38.74	54.00
<i>overall</i>	70.73	71.86

Table 3: Results with bigram word interpolation

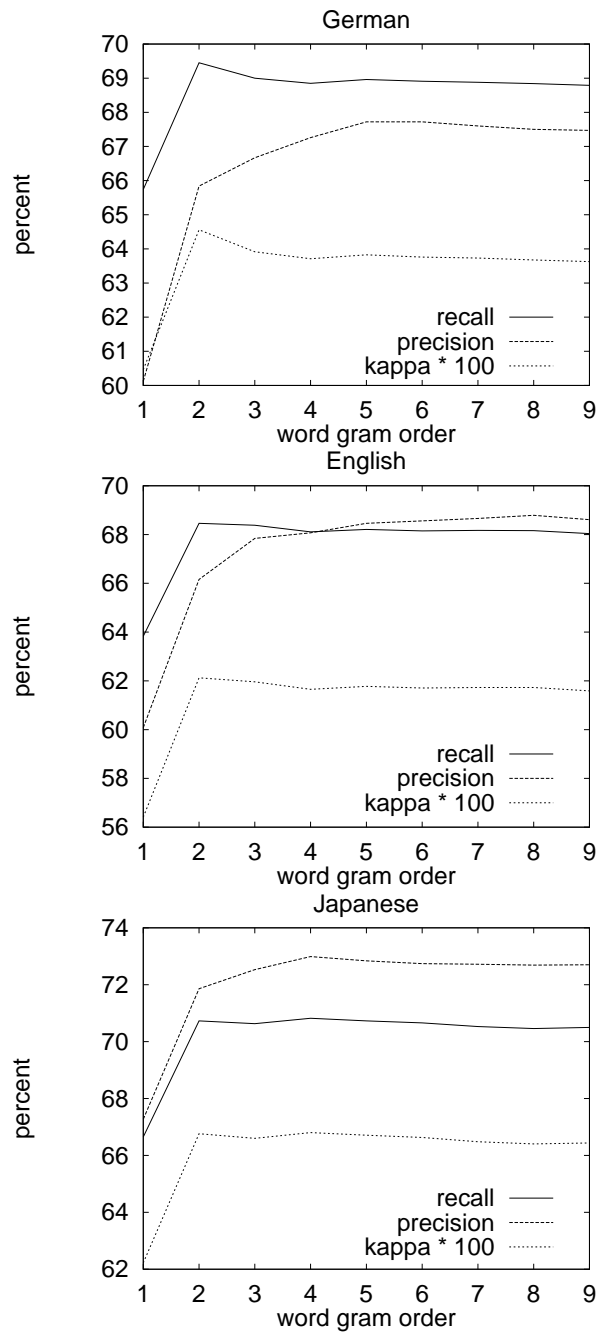


Figure 2: Recall, precision, and $\kappa * 100$ for n from 1 to 9

Short dialogues have usually a higher recall. For German, the maximum

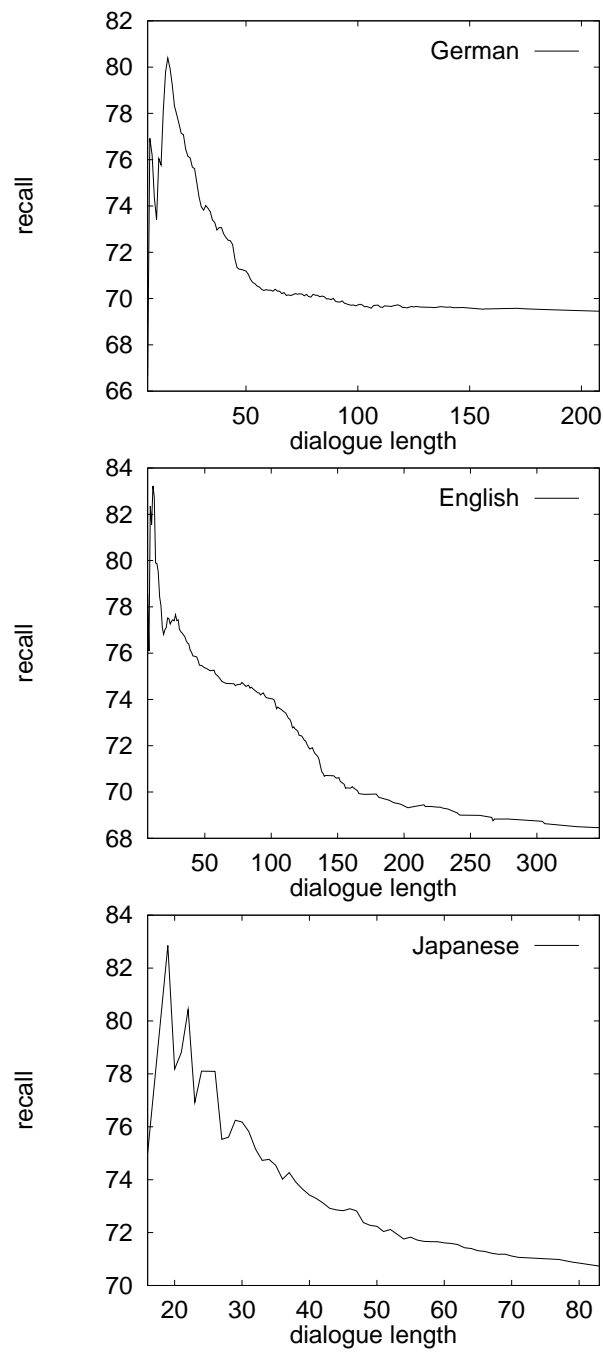


Figure 3: Cumulated recall depending on the number of acts in the dialogues

of over 80% is reached for dialogues with about 20 acts. Recall drops below 70% when the dialogue length is greater than 80 acts. For English dialogues up to the mean length of 60 acts give a recall above 75% and drop for longer dialogues. Japanese shows more variation. Recall drops below 75% with dialogues consisting of 31 dialogue acts and more.

We see various reasons for this results: short dialogues in our corpus are usually very straightforwardly centered on the tasks and domains and have less deviations. As can be seen from figure 1, the majority of the training data is relatively short for German and English. Overall recognition peaks for those dialogues, where the most and the most comparable training material is available — a fact that is to be expected. These short dialogues also don't include longer stretches of pure information exchanges. Those utterances labelled with INFORM are recognized badly (see tables 3) in all three languages, even if they are rather frequent in the corpus and therefore in the training sets.

5 Recognition Rates Using Characters

In [Nöth *et al.*, 1997], a topic detection method for a speech recognition system was presented which uses a language model approach to classify utterances in three topic classes. However, the base units are not words but codebook class sequences from the speech recognizer. For a three class problem they get a recognition rate of 80% using class trigrams.

This inspired us to a series of experiments, similar to those in the last section. Instead of words, we use the smallest unit easily available in our data, namely the characters of the words.

Figure 4 is analogous to figure 2, but the dialogue act specific language models are based on the characters of the words. For all three languages recall and precision are slightly better than the results gained with word bigrams, when using character 9-grams.

6 Comparison and Conclusions

A comparison of different approaches for dialogue act classification is currently almost impossible. The only approaches that were trained and tested on exactly the same data and tag set of 18 dialogue acts are [Reithinger and Klesen, 1997] and [Samuel *et al.*, 1998]. The first reports an overall recall value of 74.7%, the latter 73.1% for an English corpus. The neural network approach in [Kipp, 1998] performed slightly worse than a statistical n -gram model approach on the same

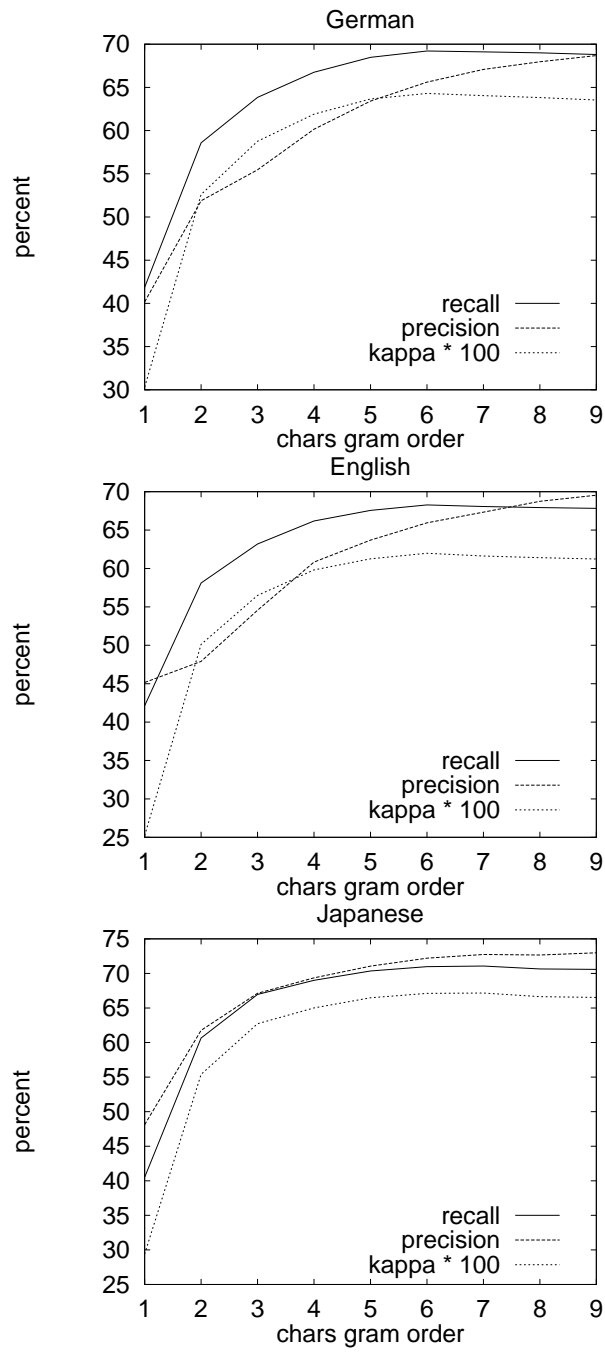


Figure 4: Recall, precision, and $\kappa * 100$ for n from 1 to 9 using characters as basis

German corpus, also using 18 dialogue acts. Other statistical approaches like [Choi *et al.*, 1999] who report a 81.6% recall and use a Korean corpus with 19 dialogue acts, differ in language, corpus size, or tag set. All of these evaluations use a fixed set of dialogues for training and testing. To our knowledge, there exists no leave on out evaluation for dialogue act recognition.

A valid comparison of methods is very difficult as long as no standardized annotation schemes and corpora are available that can be used for benchmarks. This situation is changing now: there are activities going on to define a common ground in the area of dialogue act schemes that are used to annotate corpora [Carletta *et al.*, 1997, Core *et al.*, 1999, Klein, 1999]. Even if this does not lead to one common scheme, it facilitates the transformation of annotated material from one scheme to another. Annotated corpora are already available or will be available in the near future to all interested parties. For example, the *Verbmobil* corpus is right now available from <http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html> in its major parts and will be released completely soon.

For the annotated *Verbmobil* corpus, we presented its characteristics in some detail. This is important to be able to judge learning algorithms trained on the corpus. We made various experiments to see how the n -gram order, the length of the dialogues, and the units in language models influence the classification results.

Based on the experiments presented, we optimized the dialogue act recognition module for *Verbmobil*. We use the presented subset of tags and bigram language models. The recognizer is part of one of the four translation tracks of *Verbmobil* [Reithinger, 1999, Reithinger and Engel, 2000]. Also, the version integrated in the system provides vital information for dialogue processing and dialogue summary generation [Kipp *et al.*, 2000].

References

- [Alexandersson *et al.*, 1998] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. Dialogue Acts in VERBMOBIL-2 – Second Edition. *Verbmobil-Report 226*, DFKI Saarbrücken, Universität Stuttgart, Technische Universität Berlin, Universität des Saarlandes, 1998.
- [Carletta *et al.*, 1997] Jean Carletta, Nils Dahlbäck, Norbert Reithinger, and Marilyn A. Walker, editors. *Standards for Dialogue Coding in Natural Language Processing*, Schloß Dagstuhl, 1997. Seminar Report 167.

- [Carletta, 1996] Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistics. *Computational Linguistics*, 22(2):249–254, June 1996.
- [Choi *et al.*, 1999] Won Seug Choi, Jeon-Mi-Cho, and Jungyun Seo. Analysis System of Speech Acts and Discourse Structures Using Maximum Entropy Model. In *Proceedings of Association of Computational Linguistics*, pages 230–237, Baltimore, 1999.
- [Core *et al.*, 1999] Mark Core, Masato Ishizaki, Johanna Moore, Christine Nakatani, Norbert Reithinger, David Traum, and Syun Tutiya. The Report of The Third Workshop of the Discourse Resource Initiative. Technical Report 3 (CC-TR-99-1), Chiba University, 1999.
- [Garner *et al.*, 1996] Phil Garner, Sue Browning, Roger Moore, and Martin Russell. A Theory of Word Frequencies and its Application to Dialogue Move Recognition. In *Proceedings of the International Conference on Speech and Language Processing*, pages 1880–1883, Philadelphia, PA., 1996.
- [Kipp *et al.*, 2000] Michael Kipp, Jan Alexandersson, Ralf Engel, and Norbert Reithinger. Dialog Processing. In Wolfgang Wahlster, editor, *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Springer, 2000.
- [Kipp, 1998] Michael Kipp. The Neural Path to Dialogue Acts. In *Proceedings of the European Conference on Artificial Intelligence*, pages 175–179, 1998.
- [Klein, 1999] Marion Klein. Standardisation Efforts on the Level of Dialogue Act in the MATE Project. In *Proceedings of the ACL Workshop "Towards Standards and Tools for Discourse Tagging"*, pages 35–41, 1999.
- [Klesen, 1997] Martin Klesen. Statistische Klassifikation von Dialogakten. Master's thesis, Universität des Saarlandes, Saarbrücken, 1997.
- [Mast *et al.*, 1995] Marion Mast, Heinrich Niemann, Elmar Nöth, and Ernst Günter Schukat-Talamazzini. Automatic Classification of Dialog Acts with Semantic Classification Trees and Polygrams. International Joint Conference on Artificial Intelligence, Workshop on Machine Learning, Montreal, 1995.
- [Nöth *et al.*, 1997] Elmar Nöth, Stefan Harbeck, Heinrich Niemann, and Volker Warnke. A Frame and Segment Based Approach for Topic Spotting. In *Proceedings of EuroSpeech-97*, pages 275 – 278, 1997.

- [Nöth *et al.*, 1999] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the Use of Prosody in Automatic Dialogue Understanding. In *Proc. ESCA Workshop on Dialogue and Prosody*, pages 25–34, Eindhoven, Netherlands, 1999.
- [Reithinger and Engel, 2000] Norbert Reithinger and Ralf Engel. Robust Content Extraction for Translation and Dialog Processing. In Wolfgang Wahlster, editor, *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Springer, 2000.
- [Reithinger and Klesen, 1997] N. Reithinger and M. Klesen. Dialogue Act Classification Using Language Models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, Rhodes, 1997.
- [Reithinger *et al.*, 1996] Norbert Reithinger, Ralf Engel, Michael Kipp, and Martin Klesen. Predicting Dialogue Acts for a Speech-To-Speech Translation System. In *Proceedings of the International Conference on Speech and Language Processing*, pages 654–657, Philadelphia, PA, 1996.
- [Reithinger, 1999] Norbert Reithinger. Robust information extraction in a speech translation system. In *Proceedings of EuroSpeech-99*, pages 2427–2430, 1999.
- [Samuel *et al.*, 1998] Ken Samuel, Sandra Carberry, and K. Vijay-shanker. Computing Dialogue Acts from Features with Transformation-Based Learning. In *Proceedings of the American Association for Artificial Intelligence*, pages 90–97, 1998.
- [Tanaka and Yokoo, 1999] Hideki Tanaka and Akio Yokoo. An Efficient Statistical Speech Act Type Tagging System for a Speech Translation System. In *Proceedings of the Association for Computational Linguistics*, pages 381–388, Baltimore, 1999.
- [Wahlster, 2000] Wolfgang Wahlster, editor. *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Springer, 2000.