# Learning dialog act processing

**Stefan Wermter and Matthias Löchel**
Computer Science Department
University of Hamburg
22765 Hamburg
Germany
wermter@informatik.uni-hamburg.de
löchel@informatik.uni-hamburg.de

## Abstract

In this paper we describe a new approach for *learning* dialog act processing. In this approach we integrate a symbolic semantic segmentation parser with a learning dialog act network. In order to support the unforeseeable errors and variations of spoken language we have concentrated on robust data-driven learning. This approach already compares favorably with the statistical average plausibility method, produces a segmentation and dialog act assignment for all utterances in a robust manner, and reduces knowledge engineering since it can be bootstrapped from rather small corpora. Therefore, we consider this new approach as very promising for learning dialog act processing.

## 1 Introduction

For several decades, the pragmatic interpretation at a dialog act level belongs to the most difficult and challenging tasks for natural language processing and computational linguistics (Austin, 1962; Searle, 1969; Wilks, 1985). Recently, we can see an important development in natural language processing and computational linguistics towards the use of empirical learning methods (for instance, (Charniak, 1993; Marcus et al., 1993; Wermter, 1995; Jones, 1995; Wermter et al., 1996)).

Primarily, new learning approaches have been successful for *lexically or syntactically tagged text corpora*. In this paper we want to examine the potential of learning techniques at *higher pragmatic dialog levels of spoken language*. Learning at least part of the dialog knowledge is desirable since it could reduce the knowledge engineering effort. Furthermore, inductive learning algorithms work in a data-driven mode and have the ability to extract gradual regularities in a robust manner. This robustness is particularly important for processing spoken language since spoken language can contain constructions including interjections, pauses, corrections, repetitions, false starts, semantically or syntactically incorrect constructions, etc.

The use of learning is a new approach at the level of dialog acts and only recently, there have been some learning approaches for dialog knowledge (Mast et al., 1996; Alexanderson et al., 1995; Reithinger and Maier, 1995; Wang and Waibel, 1995). Different from these approaches, in this paper we examine the combination of learning techniques in simple recurrent networks with symbolic segmentation parsing at a dialog act level.

Input to our dialog component are utterances from a corpus of business meeting arrangements like: "Tuesday at 10 is for me now again bad because I there still train I think we should [delay] the whole then really to the next week is this for you possible"[1]. For a flat level of dialog act processing, the incremental output is (1) utterance boundaries within a dialog turn and (2) the specific dialog act within an utterance. The paper is structured as follows: First we will outline the domain and task and we will illustrate the dialog act categories. Then, we will describe the overall architecture of the dialog component in the SCREEN system (Symbolic Connectionist Robust EnterprisE for Natural language), consisting of the segmentation parser and the dialog act network. We will describe the learning and generalization results for this dialog component and we will point out contributions and further work.

[1] This is almost a literal translation of the German utterance: "Dienstags um zehn ist bei mir nun wiederum schlecht weil ich da noch trainieren bin ich denke wir sollten das Ganze dann doch auf die nächste Woche verschieben geht es bei ihnen da." We have chosen the literal word-by-word translation since our processing is incremental and knowledge about the order of the German words matter for processing.

## 2 The Task

The main task is the examination of learning for dialog act processing and the domain is the arrangement of business dates. For this domain we have developed a classification of dialog acts which is shown in table 1 together with examples. Our guideline for the choice of these dialog acts was based on (1) the particular domain and corpus and (2) our goal to learn rather few dialog categories but in a robust manner[2].

| Dialog act (Abbreviation) | Example |
|---|---|
| acceptance (acc) | That would be fine |
| query (query) | Do you know Hamburg |
| rejection (rej) | This is too late for me |
| request comment (re-c) | Is that possible |
| request suggestion (re-s) | When would it be ok |
| statement (state) | Right, it's a Tuesday |
| date/loc. suggestion (sug) | I propose April 13th |
| miscellaneous (misc) | So long, bye |

Table 1: Dialog acts and examples

For example, in our example turn below there are several utterances and each of them has a particular dialog act as shown below. The turn starts with a rejection, followed by an explaining statement. Then a suggestion is made and a request for commenting on this suggestion:

- Dienstags um zehn ist bei mir nun wiederum schlecht (Tuesday at 10 is for me now again bad) → rejection
- weil ich da noch trainieren bin (because I there still train) → statement
- ich denke (I think) → miscellaneous
- wir sollten das Ganze dann doch auf die naechste Woche verschieben (we should the whole then really to the next week delay; we should delay the whole then really to the next week) → suggestion
- geht es bei ihnen da (is that for you possible) → request comment

It is important to note that segmentation parsing and dialog act processing work incremental and in parallel on the incoming stream of word hypotheses. After each incoming word the segmentation parsing and dialog act processing analyze the current input. For instance, dialog act hypotheses are available with the first input word, although good hypotheses may only be possible after most of an utterance has been seen. Our general goal here is to produce hypotheses about segmentation and dialog acts as early as possible in an incremental manner.

## 3 The Overall Approach

The research presented here is embedded in a larger effort for examining hybrid connectionist learning capabilities for the analysis of spoken language at various acoustic, syntactic, semantic and pragmatic levels. To investigate hybrid connectionist architectures for speech/language analysis we developed the SCREEN system (Symbolic Connectionist Robust EnterprisE for Natural language) (Wermter and Weber, 1996). For the task of analyzing *spontaneous* language we pursue a shallow screening analysis which uses primarily flat representations (like category sequences) wherever possible.

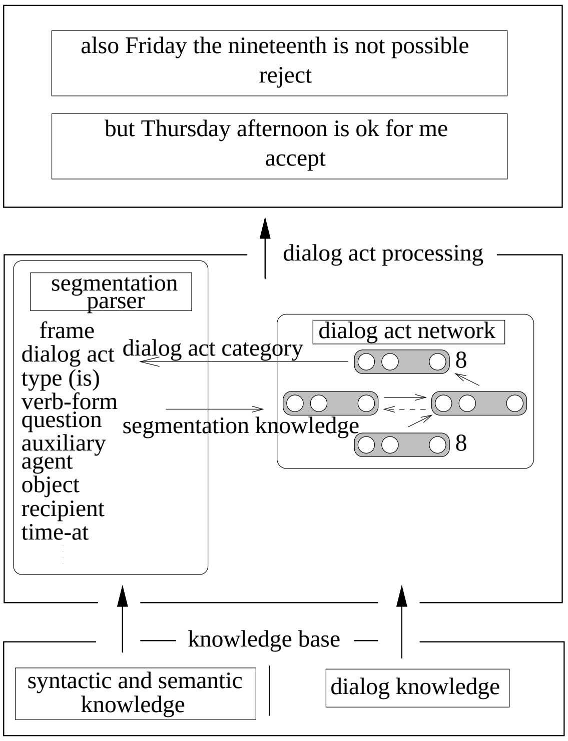

Figure 1: Architecture of dialog act component

Figure 1 gives an overview of our dialog component in SCREEN. The interpretation of utterances is based on syntactic, semantic and dialog knowledge for each word. The syntactic and semantic knowledge is provided by other SCREEN components and has been described elsewhere (Wermter and Weber, 1995). Each word of an utterance is processed incrementally and passed to the seg-

[2] This is also motivated by our additional goal of receiving noisy input directly from a speech recognizer.

mentation parser and to the dialog act network. The dialog act network provides the currently recognized dialog act for the current flat frame representation of the utterance part. The segmentation parser provides knowledge about utterance boundaries. This is important control knowledge for the dialog act network since without knowing about utterance boundaries the dialog network may assign incorrect dialog acts.

## 4 The Segmentation Parser

The segmentation parser receives one word at a time and builds up a flat frame structure in an incremental manner (see tables 2 and 3). Together with each word the segmentation parser receives syntactic and semantic knowledge about this word based on other syntactic and semantic modules in SCREEN. Each word is associated with 1. its most plausible basic syntactic category (e.g. noun, verb, adjective), 2. its most plausible abstract syntactic category (e.g. noun group, verb group, prepositional group), 3. basic semantic category (e.g., animate, abstract), and 4. abstract semantic category (e.g., agent, object, recipient).

| Slots | 3. Phrase | Final Phrase |
|---|---|---|
| dialog act | *cat?* | *reject* |
| type | is | is |
| verb-form | ((is)) | ((is)) |
| question | nil | nil |
| auxiliary | nil | nil |
| agent | nil | nil |
| object | nil | nil |
| recipient | | ((for me)) |
| time-at | ((Tuesday) (at 10)) | ((Tuesday) (at 10)) |
| time-from | nil | nil |
| time-to | nil | nil |
| location-at | nil | nil |
| location-from | nil | nil |
| location-to | nil | nil |
| confirm | nil | nil |
| negation | nil | ((bad)) |
| miscellaneous | nil | ((now again)) |
| input | Tuesday at 10 is | Tuesday at 10 is for me now again bad |

Table 2: Incremental slot filling in frame 1: literal incremental translation: Dienstags um zehn ist bei mir nun wiederum schlecht (Tuesday at 10 is for me now again bad)

This syntactic and semantic category knowledge is used by the segmentation parser for two main purposes. First, this category knowledge is needed for our segmentation heuristics. For our domain we have developed segmentation rules which allow the system to split turns into utterances. For instance, if we know that the basic syntactic category of a word "because" is conjunction and it is part of a conjunction group, then this is an indication to close the current frame and trigger a new frame for the next utterance. Second, the category knowledge, primarily the abstract semantic knowledge, is used for filling the frames, so that we get a symbolically accessible structure rather than a tagged word sequence.

| Slots | 1.–3. Phrase | Final Phrase |
|---|---|---|
| dialog act | *cat?* | *statement* |
| type | move | move |
| verb-form | nil | ((train)) |
| question | nil | nil |
| auxiliary | nil | am |
| agent | ((I)) | ((I)) |
| object | nil | nil |
| recipient | nil | nil |
| time-at | nil | nil |
| time-from | nil | nil |
| time-to | nil | nil |
| location-at | nil | nil |
| location-from | nil | nil |
| location-to | nil | nil |
| confirm | nil | nil |
| negation | nil | nil |
| miscellaneous | ((because) (there still)) | ((because) (there still)) |
| input | because I there still | because I there still train am |

Table 3: Incremental slot filling in frame 2;...weil ich da noch trainieren bin (because I there still train [am])

The segmentation parser is able to segment 84% of the 184 turns with 314 utterances correctly. The remaining 16% are mostly difficult ambiguous cases some of which could be resolved if more knowledge could be used. For instance, while many conjunctions like "because" are good indicators for utterance borders, some conjunctions like "and" and "or" may not start new coordinated subsentences but coordinate noun groups. Fundamental structural disambiguation could be used to deal with these cases. Since they occur relatively rarely in our spoken utterances we have chosen not to incorporate structural disambiguation. Furthermore, another class of errors is characterized by time and location specifiers which can occur at the end or start of an utterance. For instance, consider the example: "On Tuesday the sixth of April I still have a slot in the afternoon — is that possible" versus "On Tuesday the sixth of April I still have a slot — in the afternoon is that possible". Such decisions are difficult and additional knowledge like prosody might help here. Currently, there is a preference for filling the earlier frame.

## 5 The Dialog Act Network

In table 1 we have described the dialog acts we use in our domain. Before we start to describe any experiments on learning dialog acts we show the distribution of dialog acts across our training and test sets. Table 4 shows the distribution for our set of 184 turns with 314 utterances. There were 100 utterances in the training set and 214 in the test set. As we can see, suggestions and explanatory statements often occur but in general all dialog acts occur reasonably often. This distribution analysis is important for judging the learning and generalization behavior.

| Category | Training | Test |
|---|---|---|
| sug | 31% | 26% |
| state | 20% | 21% |
| rej | 12% | 10% |
| misc | 11% | 18% |
| re-s | 10% | 8% |
| acc | 9% | 12% |
| query | 5% | 3% |
| re-c | 2% | 3% |

Table 4: Distribution of the dialog acts in training and test set

After this initial distribution analysis we now describe our network architecture for learning dialog acts. Dialog acts depend a lot on significant words and word order. Certain key words are much more significant for a certain dialog act than others. For instance "propose" is highly significant for the dialog act *suggest*, while "in" is not. Therefore we computed a smoothed dialog act plausibility vector for each word $w$ which reflects the plausibility of the categories for a particular word. The sum of all values is 1 and each value is at least 0.01. The plausibility value of a word $w$ in a dialog category $da_i$ with the frequency $f$ is computed as described in the formula below.

$$\frac{f_{da_i}(w) - (f_{da_i}(w) * f_{da_j} = {}_0(da_j) * 0.01)}{Total\ frequency\ f(w)\ in\ corpus}$$

Table 5 shows examples of plausibility vectors for some words. As we can see, "bad" has the highest plausibility for the *reject* dialog act, and "propose" for the *suggest* dialog act. On the other hand the word "is" is not particularly significant for certain dialog acts and therefore has a plausibility vector with relatively evenly distributed values.

|  | bad | propose | is |
|---|---|---|---|
| acc | 0.28 | 0.01 | 0.22 |
| misc | 0.01 | 0.38 | 0.02 |
| query | 0.01 | 0.01 | 0.07 |
| rej | 0.66 | 0.01 | 0.34 |
| re-c | 0.01 | 0.01 | 0.01 |
| re-s | 0.01 | 0.01 | 0.02 |
| state | 0.01 | 0.01 | 0.27 |
| sug | 0.01 | 0.56 | 0.05 |

Table 5: Three examples for plausibility vectors

We have experimented with different variations of simple recurrent networks (Elman, 1990) for learning dialog act assignment. We had chosen simple recurrent networks since these networks can represent the previous context in an utterance in their recurrent context layer. The best performing network is shown in figure 2.

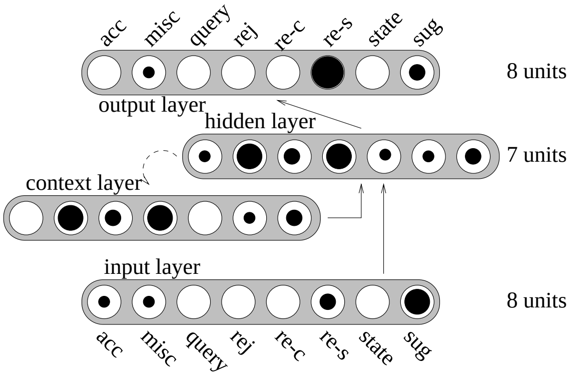

Figure 2: Dialog act network with dialog plausibility vectors as input

Input to this network is the current word represented by its dialog plausibility vector. The output is the dialog act of the whole utterance. Between input and output layer there are the hidden layer and the context layer. All the feedforward connections in the network are fully connected. Only the recurrent connections from the hidden layer to the context layer are 1:1 copy connections, which represent the internal learned context of the utterance before the current word. Training in these networks is performed by using gradient descent (Rumelhart et al., 1986) using up to 3000 cycles through the training set. By using the internal learned context it is possible to make dialog act assignments for a whole utterance. While processing a whole utterance, each word is presented with its plausibility vector and at the output layer we can check the incrementally assigned dialog acts for each incoming word of the utterance.

We have experimented with different input knowledge (only dialog act plausibility vectors, additional abstract semantic plausibility vectors, etc.), different architectures (different numbers of context layers, and different number of units in hidden layer, etc). Due to space restrictions it is not possible to describe all these comparisons. Therefore we just focus on the description of the network with the best generalization performance.

| Dialog acts | Training | Test |
|---|---|---|
| acc | 88.9 | 72.0 |
| state | 90.0 | 90.9 |
| misc | 54.5 | 73.7 |
| query | 40.0 | 0.0 |
| rej | 91.7 | 85.7 |
| sug | 90.3 | 92.9 |
| re-c | 0.0 | 0.0 |
| re-s | 90.0 | 82.4 |
| Total | 82.0 | 79.4 |

Table 6: Performance of simple recurrent network with dialog plausibility vectors in percent

Table 6 shows the results for our training and test utterances. The overall performance on the training set was 82.0% on the training set and 79.4% on the test set. An utterance was counted as classified in the correct dialog act class if the majority of the outputs of the dialog act network corresponded with the desired dialog act. This good performance is partly due to the distributed representation in the dialog plausibility vector at the input layer. Other second best networks with additional local representations for abstract semantic category knowledge could perform better on the training set but failed to generalize on the test set and only reached 71%.

The remaining errors are partly due to seldomly occurring dialog acts. For instance, there are only 2% of the training utterances and 2.8% of the test utterances which belong to the request-comment dialog act. The network was not able to learn correct assignments due to the little training data. The drop in the performance for the query dialog act from training to test set can be explained by the higher variability of the queries compared to all other categories. Since queries differ much more from each other than all other dialog acts they could not be generalized. However they do not occur very often. All other often occurring dialog act categories performed very well as the individual percentages and the overall percentage show.

## 6 Discussion and Conclusions

What do we learn from this? When we started this work it was not clear to what extent a symbolic segmentation parser and a connectionist learning dialog act network could be integrated to perform an analysis at the semantics and dialog level. We have shown that a symbolic segmentation parser and a learning dialog network can be integrated to perform dialog act assignments for spoken utterances. While other related work has focused on statistical learning we have explored the use of learning in simple recurrent networks. Our corpus of 2228 words is still medium size. Nevertheless, we consider the results as promising, given that it is - to the best of our knowledge - the first attempt to integrate symbolic segmentation parsing with dialog act learning in simple recurrent networks.

How well do we perform compared to related work? In spite of many projects in the ATIS and VERBMOBIL domains there is not a lot of work on *learning* for the dialog level. However, recently there have been some investigations of statistical techniques (Reithinger and Maier, 1995) (Alexanderson et al., 1995) (Mast et al., 1996). For instance Mast and colleagues report 58% for learning dialog act assignment with semantic classification trees and 69% for learning with pentagrams but they also used more categories than in our approach so that the approaches are not directly comparable.

For a further evaluation of our trained network architecture we compared our results with a statistical approach based on the same data. Plausibility vectors for dialog acts represent the distribution of dialog acts for each word for the current corpus. However, for assigning a dialog act to a whole utterance all the words of this utterance have to be considered. A simple but efficient approach would be to compute the average plausibility vector for each utterance which has been found. Then the dialog act with the highest averaged plausibility vector for a complete utterance would be taken as the computed dialog act. This statistical approach reached a performance of 62% correctness on the training and test set compared to the 82% and 79% of our dialog network. So simple recurrent networks performed better than the statistical average plausibility method. In comparison to statistical techniques which have also been used successfully on large corpora, it is our understanding that simple recurrent networks may be particularly suitable for domains where only smaller corpora are available or where classification data is hard to get (as it is the case for pragmatic dialog acts.)

What will be further work? So far we have concentrated on single utterances and we do not account for the relationship between utterances in a dialog. While we could demonstrate that such a local strategy could assign correct dialog acts in many cases, it might be interesting to explore to what extent knowledge about previous dialog acts in previous utterances could even improve our results. Furthermore, we have developed the segmentation parser and dialog act network as very robust components. In fact, both are very robust in the sense that they will *always* produce the best possible segmentation and dialog act categorization. In the future we plan to explore how the output from a speech recognizer can be processed by our dialog component. Sentence and word hypotheses from a speech recognizer are still far from optimal for continuously spoken spontaneous speech. Therefore we have to account for highly ungrammatical constructions. The segmentation parser and the dialog network already contain the robustness which is a precondition for dealing with real-world speech input.

## Acknowledgements

This research was funded by the German Federal Ministry for Research and Technology (BMBF) under Grant #01IV101A0 and by the German Research Association (DFG) under contract DFG Ha 1026/6-2. We would like to thank S. Haack, M. Meurer, U. Sauerland, M. Schrattenholzer, and V. Weber for their work on SCREEN.

## References

J. Alexanderson, E. Maier, and N. Reithinger. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *Proceedings of the European Association for Computational Linguistics*, Dublin.

J. Austin. 1962. *How to do things with words.* Clarendon Press, Oxford.

E. Charniak. 1993. *Statistical Language Learning.* MIT Press, Cambridge, MA.

J. L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–221.

D. Jones, editor. 1995. *New Methods in Language Processing.* University College London.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(1).

M. Mast, E. Noeth, H. Niemann, and E. G. Schukat Talamazzini. 1996. Automatic classification of dialog acts with semantic classification trees and polygrams. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 217–229. Springer, Heidelberg.

N. Reithinger and E. Maier. 1995. Utilizing statistical dialogue act processing in verbmobil. In *Computational Linguistics Archive.*

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 318–362. MIT Press, Cambridge, MA.

J. R. Searle. 1969. *Speech Acts.* Cambridge University Press, Cambridge.

Y. Wang and A. Waibel. 1995. Connectionist transfer in machine translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 37–44, Tzigov Chark.

S. Wermter and V. Weber. 1995. Artificial neural networks for automatic knowledge acquisition in multiple real-world language domains. In *Proceedings of the International Conference on Neural Networks and their Applications*, Marseille.

S. Wermter and V. Weber. 1996. Interactive spoken language processing in the hybrid connectionist system SCREEN: learning robustness in the real world. *IEEE Computer*, 1996. in press.

S. Wermter, E. Riloff, and G. Scheler. 1996. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing.* Springer, Berlin.

S. Wermter. 1995. *Hybrid Connectionist Natural Language Processing.* Chapman and Hall, London, UK.

Y. Wilks. 1985. Relevance, points of view and speech acts: An artificial intelligence view. Technical Report MCCS-85-25, New Mexico State University.