EVALUATION OF DIALOGUE ACT RECOGNITION APPROACHES

Pavel Král[1], *Tomáš Pavelka*[1*]

[1]Dept. Informatics & Computer Science
University of West Bohemia
Plzeň, Czech Republic
{pkral,tpavelka}@kiv.zcu.cz

Christophe Cerisara[2]

[2]LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
France
cerisara@loria.fr

ABSTRACT

This paper deals with automatic dialogue act recognition. Dialogue acts (DAs) are utterance-level labels that represent different states of a dialogue, such as questions, statements, hesitations, etc. Information about actual DA can be seen as the first level of dialogue understanding. The main goal of this paper is to compare our dialogue act recognition approaches that model the utterance structure, and are particularly useful when the DA corpus is small, with n-gram based approaches. Our best approach is also combined successfully with prosodic models. We further show that sentence structure-based approaches significantly outperform n-gram based methods.

1. INTRODUCTION

Modeling and automatically identifying the spontaneous dialogue structure is very important in order to better interpret and understand speech. What should actually be modeled is still an open issue, but several specific characteristics of dialogue have already been clearly identified. Dialogue Acts (DAs) are one of these characteristics.

Austin defines in [1] the dialogue act as the meaning of an utterance at the level of illocutionary force. In other words, the dialogue act is the function of a sentence (or its part) in the dialogue. For example, the function of a question is to request some information, while an answer shall provide this information.

The DA recognition module shall be used to improve the performance of an automatic dialogue system by allowing it to better interpret the user input. It can also be integrated into an automatic speech recognizer to improve language modeling, e.g. by choosing a DA dependent language model.

In automatic DA recognition, lexical and syntactic information is often modeled by probabilistic n-gram models. However, these n-grams usually represent local structures only. Conceiving general grammars is still an open issue, especially for spontaneous speech.

We proposed in [2, 3, 4] to include a simplified information related to the utterance structure, i.e. the position of the words within the utterance. This method presents the advantage of introducing valuable information related to the global utterance structure, without increasing the complexity of the overall system. We have shown that the DA recognition accuracy increases when utterance structure information is used.

We now extend our work by proposing new utterance structure models, and by comparing the performance of our approaches with several n-gram based methods. Finally, our best method is further combined with prosodic model. All methods are evaluated on a small Czech DA corpus.

This paper is organized as follows. Section 2 presents some related works about automatic dialogue act recognition. Section 3 describes the different models we propose. Section 4 gives experimental results for our methods. In the last section, we discuss the research results and we propose some future research directions.

2. RELATED WORK

To the best of our knowledge, few studies on dialogue act modeling and automatic recognition have been published for the Czech language. Conversely, there are several work for other languages, especially for English and German.

Different sets of dialogue acts are defined in these works, depending on the target application and the available corpora. In [5], 42 dialogue acts classes are defined for English, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [6]. Switchboard-DAMSL tag-set [7] (SWBD-DAMSL) is an adaptation of DAMSL in the domain of telephone conversation. The Meeting Recorder DA (MRDA) tag-set [8] is another very popular tag-set, which is based on the SWBD-DAMSL taxonomy. MRDA contains 11 general DA labels and 39 specific labels. Jekat [9] defines for German and Japanese 42 DAs, with 18

*This work has been partly supported by the Ministry of Education, Youth and Sports of Czech Republic grant (NPV II-2C06009).

DAs at the illocutionary level, in the context of the VERBMOBIL corpus. The Map-Task [10] is another English tagset. It contains 19 DA tags that are structured into three levels.

These complete DA tag-sets are usually reduced for recognition into a few broad classes, because some classes occur rarely, or because other DAs are not useful for the target application. One typical regrouping may be [11]:

- statements
- questions
- backchannels
- incomplete utterance
- agreements
- appreciations
- other

Automatic recognition of dialogue acts is usually achieved using one of, or a combination of the following types of information:

1. lexical (and syntactic) information
2. prosodic information
3. context of each dialogue act

Lexical information (i.e. word sequence in the utterance) is useful for automatic DA recognition, because different DAs are usually composed from different word sequences. Some cue words and phrases can thus serve as explicit indicators of dialogue structure. For example, 88.4 % of the trigrams "<start> do you" occur in English in *yes/no questions* [12].

Several models are used to represent lexical information. Bayesian approaches such as n-gram language models [5], [13] can be used. Non-Bayesian approaches are also popular such as semantic classification trees [13], memory-based learning [14], or transformation-based learning [15].

Syntactic information is related to the *order* of the words in the utterance. For instance, in French and Czech, the relative order of the *subject* and *verb* occurrences might be used to discriminate between declarations and questions.

Words n-grams are often used to model some local syntactic information. Král et al. propose in [4] to represent word position in the utterance in order to take into account global syntactic information. Another type of syntactic information recently used for DA recognition are "cue phrases". These can be modeled with a subset of specific n-grams, where n may vary from 1 to 4, which are selected based on their capacity to predict a specific DA and on their occurrence frequency [16].

Prosodic information [11], particularly the melody of the utterance, is often used to provide additional clues to classify sentences in terms of DAs. For instance, some dialogue acts can be generally characterized by prosody as follows [17]:

- a falling intonation for statements
- a rising F0 contour for some questions (particularly for declaratives and yes/no questions)
- a continuation-rising F0 contour characterizes a (prosodic) clause boundaries, which is different from the end of utterance
- accepts have usually a higher energy, a greater F0 movement than backchannels

The following prosodic features and classifiers are further used. In [11], the duration, pause, fundamental frequency (F0), energy and speaking rate prosodic features are modeled by a CART-style decision trees classifier. In [18], prosody is used to segment utterance. The duration, pause, F0-contour and energy features are used in [19, 20]. In both [19] and [20], several features are computed based on these basic prosodic attributes, for example the max, min, mean and standard deviation of F0, the mean and standard deviation of the energy, the number of frames in utterance and the number of voiced frames. The features are computed on the whole sentence and also on the last 200 ms of each sentence. The authors conclude that the end of sentences carry the most important prosodic information for DAs recognition. Furthermore, three different classifiers, hidden Markov models, classification and regression trees and neural networks, are compared and give similar DAs recognition accuracy.

Shriberg et al. show in [11] that it is better to use prosody for DA recognition in three separate tasks, namely question detection, incomplete utterance detection and agreements detection, rather than for detecting all DAs in one task.

The dialogue act context is used to predict the most probable next dialogue acts. This context is often called "dialogue history" and can be modeled by Hidden Markov Models (HMMs) [5], Bayesian Networks [21], Discriminative Dynamic Bayesian Networks (DBNs) [22], or n-gram language models [23].

Lexical and prosodic models are most often combined in the following way [5]:

$$\begin{aligned} P(W,F|C) &= P(W|C).P(F|W,C) \\ &\simeq P(W|C).P(F|C) \end{aligned} \quad (1)$$

where C represents a dialogue act and W and F respectively represent lexical and prosodic information (assumed independent).

3. DIALOGUE ACT RECOGNITION APPROACHES

The approaches described next are based on Bayesian models. The main objective is to compute the probability that an

utterance belongs to a given DA class, given the lexical and syntactic information, i.e. the word sequence.

3.1. N-gram Language Models

N-gram DA models are quite common in the domain[11, 5] and will thus constitute our baseline model.

Let W be the word sequence in the pronounced utterance, let C be the DA class, then the recognized class is given by:

$$\begin{aligned} \hat{C} &= \arg\max_C P(C|W) \\ &= \arg\max_C P(C).P(W|C) \end{aligned} \quad (2)$$

The simplest model, unigram, assumes independence between successive words. More complex ones, such as 2-grams, 3-grams, etc., consider syntactic information about the dependencies between adjacent words. These n-grams usually model local utterance structures only.

In the following experiments, we have tested different kinds of n-gram models:

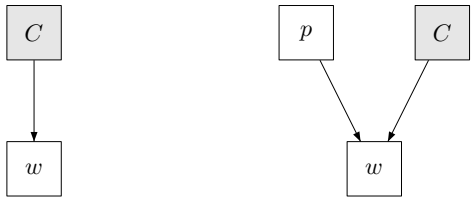
- Simple 1-gram, 2-grams and 3-grams (with standard backoff);
- Interpolated n-grams between 2-gram and 1-gram;
- Interpolated n-grams between 3-gram, 2-gram and 1-gram;

Interpolation weights have been trained on a development corpus.

3.2. Approaches exploiting utterance structure

In the following, we assume that each word is independent on the other words, but is dependent on its position in the utterance, which is modeled by a random variable p.

We can model our approach by a very simple Bayesian network with three variables, as shown in Figure 1. In this figure, C encodes the dialogue act class of the test utterance, w represents a word and p its position in the utterance.



(a) (b)

Fig. 1. Graphical model of our approaches: grayed nodes are hidden

In the left model of Figure 1, $P(w|C,p)$ is assumed independent of the position: $P(w|C,p) \simeq P(w|C)$. This system only considers lexical information, and the probability over the whole utterance is given by equation 3.

$$P(w_1, \cdots, w_T|C) = \prod_{i=1}^{T} P(w_i|C) \quad (3)$$

This model corresponds to unigrams.

On the right part of Figure 1, information about the position of each word is included. However, this model may not be used directly, because the new variable p greatly reduces the ratio between the size of the corpus and the number of free parameters to train.

We have proposed in [2, 4] three methods to solve this problem. These methods are described next. Then, two new approaches are proposed: interpolated multiscale position and frequency bin interpolation.

3.2.1. Multiscale Position

This method exploits a description of the utterance in several levels to smooth the probabilities across these levels. Random variable p can take a different number of values depending on the scale. All these scales are represented in a dyadic tree. During training, n-gram models are trained, starting at the broadest scale and going down the tree to the leaves: when there is not enough occurrences to reliably compute the parameters, the model from the upper level is copied down. Classification is then realized at the finest scale based on the following equation:

$$\begin{aligned} \hat{C} &= \arg\max_C P(C|w_1, \cdots, w_T, p_1, \cdots, p_T) \\ &= \arg\max_C P(C) \prod_{i=1}^{T} P(w_i|C, p_i) \end{aligned} \quad (4)$$

3.2.2. Non-linear Merging

This method encodes dependency between W and p by a non-linear function that includes p. A muti-layer Perceptron (MLP) is used for this purpose. The recognized class is given by:

$$\hat{C} = \arg\max_C P(C|w_1, \cdots, w_T, p_1, \cdots, p_T) \quad (5)$$

3.2.3. Best Position Approach

The random variable p now represents the best position amongst every possible position, i.e. the position that minimizes the DA recognition error rate. It is then possible to recognize DAs by the following equation [4]:

$$\hat{C} = \arg\max_C P(C) \sum_p P(w_p|C)P(p|C) \quad (6)$$

where w_p is the word of the actual utterance at the best position p. The lexical likelihood $\prod_i P(w_i|C)$ used previously is now replaced by the weighted sum of each word likelihood, where weights represent the importance of each position.

Compared to the two previously proposed approaches, this alternative presents the advantage of decoupling the position model from the lexical model. The lexical models $P(w_i|C)$ are thus still trained on the whole corpus, which is not divided into position-relative clusters as in the multiscale approach.

3.2.4. Interpolated Multiscale Position

The original multiscale position method uses a simple backoff scheme to choose among different levels: If the count of the currently processed word for the given position, in the given level is below a chosen threshold (i.e. the model is poorly trained) the model is replaced by one from the upper level. For example if the current level has 8 positions (and thus 8 unigram models) and the count for the current word is below the threshold, the model from the level above the current one (which in our case has 4 positions) is used. This is done recursively until either a model with sufficient count is found, or the upper most level (a single unigram) is reached.

We have tried an alternative method for smoothing probabilities across levels, which is based on linear interpolation (see e.g. [24]). When computing the probability of a given word w_i the result is given as a linear interpolation of unigram probabilities at different levels l:

$$P(w_i|C) = \sum_{l=1}^{N} \lambda_l P\big(w_i|C, p_l(i)\big) \qquad (7)$$

where positions $p_l(i)$ are computed independently at each level. The weights λ_l are trained by the Expectation Maximization (EM) algorithm on a development corpus.

3.2.5. Frequency Bin Interpolation

The "frequency bin interpolation" is an extension of the interpolated multiscale position where several weights are computed per level.

In the previous approach, the weights are trained to globally compensate for poorly trained models at a given level. A single weight is applied to the whole level regardless of whether the model at the current position is poorly trained or not.

We now propose to cluster all the words at a given level into several classes, called frequency bins, depending on their number of occurrences. Different weights are then assigned to distinct frequency bins. With this method the weights can take into account whether the unigram model at the given level and position is sufficiently trained.

3.3. Combination with Prosody

Only the two most important prosodic features as suggested in [25] are used: F0 and energy. Let us call F the set of prosodic features for one utterance. We use a Gaussian Mixture Model (GMM) classifier that computes $P(F|C)$. The best DA class is then:

$$\hat{C} = \arg\max_C P(C|F) = \arg\max_C P(F|C)P(C) \qquad (8)$$

Our prosodic approach is described in details in [26].

The outputs of the lexical, position and prosodic model are then normalized in order to obtain $P(C|W)$, $P(C|W,P)$ and $P(C|F)$. These probabilities are then combined with a Multi-Layer Perceptron (MLP), as described in [26].

4. EVALUATION

4.1. Dialogue Act Corpus

The Czech Railways corpus, which contains human-human dialogs, is used to validate the evaluated approaches. This corpus has been labelled manually with the following DAs: statements (s), orders (o), yes/no questions (qy) and other questions (q). The number of DAs of this corpus is shown in the second column of Table 1.

All experiments are realized using a cross-validation procedure, where 10% of the corpus is reserved for the test, and another 10% for the development set. The resulting global accuracy has a confidence interval $\pm$ 1%.

DA	#	Example	English translation
s	566	Chtěl bych jet do Písku.	I would like to go to Písek.
o	125	Najdi další vlak do Plzně!	Look for the next train to Plzeň!
qy	282	Řekl byste nám další spojení?	Can you tell me the next connection?
q	1200	Jak se dostanu do Šumperka?	How can I get to Šumperk?
Tot.	2173		

Table 1. Composition of the Czech Railways corpus

4.2. N-gram Language Models Experiments

Three language models, unigram, 2-gram and 3-gram, have been implemented in two different versions. The first version corresponds to classical n-grams with the standard back-

off technique. The second version is the "interpolated" n-gram model [24], as described in section 3.1. The corresponding DA recognition accuracies are shown in Table 2.

	accuracy in [%]				
Approach/ Classifier	s	o	qy	q	Global
N-gram models					
Unigram	93.5	77.6	96.5	89.9	**91.0**
2-gram	83.8	67.5	87.7	80.0	**84.6**
3-gram	72.9	78.3	65.2	64.3	**67.8**
Interp. 2-gram	86.4	70.8	83.6	85.4	**83.8**
Interp. 3-gram	83.8	70.0	83.2	81.8	**82.4**

Table 2. Dialogue acts recognition accuracy for different n-gram based approaches

Two conclusions can be drawn from this experiment. First, 3-grams perform worse than 2-grams, which are also worse than 1-gram. This seems in contradiction with the common knowledge that models are better when they include context. However, there is a simple explanation in our case, as the training corpus is too small to reliably train contextual n-grams: the models fit the training data but are not able to generalize correctly. This observation is confirmed by the second remark, which is that interpolated n-grams perform better than classical n-grams, thanks to the linear weights that compensate for this overtraining. Yet, interpolated n-grams do not reach the unigram performances, because interpolated n-grams do not exploit back-off technique. Hence, this experiment shows that, for DA n-gram models with a small training corpus, back-off smoothing should be preferred over interpolation.

4.3. Utterance Structure Experiments

Figure 2 shows the DA recognition accuracy of the *Multiscale position* approach when the maximum depth of the tree increases. In all cases, the value of the pruning threshold in this tree is set to 50, which has been found experimentally in [2].

The optimal depth of the tree is 3, which corresponds to 8 segments. Choosing a deeper tree is useless, as the recognition accuracy is almost constant. With a larger DA corpus, deeper trees could be used, which shall result in a better recognition accuracy.

The experimental results of the utterance structure methods described above are shown in the first section of Table 3.

The interpolated multiscale gives the lowest accuracy, which might be due to the fact that interpolation weights are global. Indeed, when splitting a weight into several frequency bins, results clearly increase. This experiment suggests that the back-off technique gives better results than

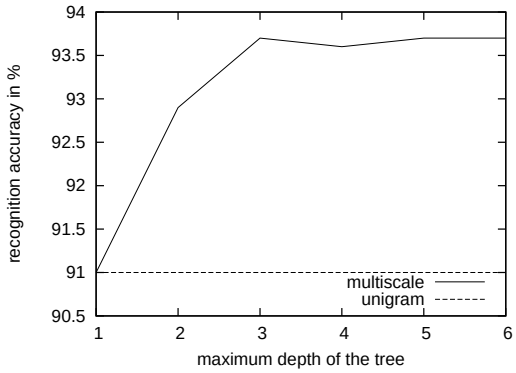


Fig. 2. Dialogue acts recognition accuracy of the multiscale position model. The x-axis represents the maximum depth of the tree, while the Y-axis shows the DA recognition accuracy

global interpolation in our multiscale approach, even though we plan next to investigate still finer interpolation weights.

When considering lexical and some syntactic information only, the best performance is obtained with the *best position* approach.

4.4. Prosody and Combination

The second section of Table 3 shows the recognition accuracy of the prosodic GMM. This recognition accuracy is obtained with a 3-mixtures GMM.

The last line of Table 3 shows the results of the combined prosodic GMM and Best position model with an MLP. The combined models gives better results than any model taken individually, which confirms that different sources of information bring different important clues to classify DAs.

	accuracy in [%]				
Approach/ Classifier	s	o	qy	q	Global
1. Utterance structure					
Multiscale	94.7	70.4	96.1	95.3	**93.8**
Non-linear	90.3	83.2	91.1	98.8	**94.7**
Best position	93.6	95.2	97.2	94.3	**95.8**
Interp. Multiscale	94.1	65.0	67.4	86.4	**76.7**
Frequency Bin	93.9	70.0	91.4	93.2	**91.1**
2. Prosodic approach					
GMM	47.7	43.2	40.8	44.3	**44.7**
3. Combination					
MLP	94.0	95.6	97.0	95.2	**96.9**

Table 3. Dialogue acts recognition accuracy for utterance structure approaches/classifiers and combination of the best approach with prosody

5. CONCLUSIONS

In this paper, several dialogue act recognition approaches have been proposed and evaluated on a small Czech corpus. We focus on the comparison of our approaches that consider utterances structure with several n-gram based methods. Experimental results show that our methods give significantly better accuracy than n-grams on this DA corpus. We also show that the combination of sentence structure-based models with prosodic information slightly increases the DA recognition accuracy.

The proposed dialogue act recognition approaches are task independent. Our perspective for the near future is to evaluate them on a larger corpus, another language and with more dialogue acts. We assume that n-gram based approaches will perform better on a larger DA corpus. Another extension of this work might thus be to combine our sentence structure approaches with n-grams and to evaluate them on several corpora. Finally, the proposed approaches might be improved by including information about the dialogue history.

6. REFERENCES

[1] J. L. Austin, "How to do Things with Words," *Clarendon Press, Oxford*, 1962.

[2] P. Král, C. Cerisara, and J. Klečková, "Automatic Dialog Acts Recognition based on Sentence Structure," in *ICASSP'06*, Toulouse, France, May 2006, pp. 61–64.

[3] P. Král, J. Klečková, T. Pavelka, and C. Cerisara, "Sentence Structure for Dialog Act recognition in Czech," in *ICTTA'06*, Damascus, Syria, April 2006.

[4] P. Král, C. Cerisara, and J. Klečková, "Lexical Structure for Dialogue Act Recognition," *Journal of Multimedia (JMM)*, vol. 2, no. 3, pp. 1–8, June 2007.

[5] A. Stolcke *et al.*, "Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech," in *Computational Linguistics*, 2000, vol. 26, pp. 339–373.

[6] J. Allen and M. Core, "Draft of Damsl: Dialog Act Markup in Several Layers," in *http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html*, 1997.

[7] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation (Coders Manual, Draft 13)," Tech. Rep. 97-01, University of Colorado, Institute of Cognitive Science, 1997.

[8] R. Dhillon, Bhagat S., H. Carvey, and Shriberg E., "Meeting Recorder Project: Dialog Act Labeling Guide," Tech. Rep. TR-04-002, International Computer Science Institute, February 9 2004.

[9] S. Jekat *et al.*, "Dialogue Acts in VERBMOBIL," in *Verbmobil Report 65*, 1995.

[10] J. Carletta, A. Isard, S. Isard, J. Kowtko, A. Newlands, G. Doherty-Sneddon, and A. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, pp. 13–31, 1997.

[11] E. Shriberg *et al.*, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?," in *Language and Speech*, 1998, vol. 41, pp. 439–487.

[12] D. Jurafsky *et al.*, "Automatic Detection of Discourse Structure for Speech Recognition and Understanding," in *IEEE Workshop on Speech Recognition and Understanding*, Santa Barbara, 1997.

[13] M. Mast *et al.*, "Automatic Classification of Dialog Acts with Semantic Classification Trees and Polygrams," in *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, 1996, pp. 217–229.

[14] M. Rotaru, "Dialog Act Tagging using Memory-Based Learning," Tech. Rep., University of Pittsburgh, Spring 2002, Term Project in. Dialog Systems.

[15] K. Samuel, S. Carberry, and K. Vijay-Shanker, "Dialogue Act Tagging with Transformation-Based Learning," in *17th international conference on Computational linguistics*, Montreal, Quebec, Canada, 10-14 August 1998, vol. 2, pp. 1150–1156, Association for Computational Linguistics, Morristown, NJ, USA.

[16] N. Webb, M. Hepple, and Y. Wilks, "Dialog act classification based on intra-utterance features," Tech. Rep. CS-05-01, Dept of Comp. Science, University of Sheffield, 2005.

[17] R. Kompe, *Prosody in Speech Understanding Systems*, Springer-Verlag, 1997.

[18] M. Mast, R. Kompe, S. Harbeck, A. Kiessling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke., "Dialog Act Classification with the Help of Prosody," in *ICSLP'96*, Philadelphia, USA, 1996.

[19] H. Wright, "Automatic Utterance Type Detection Using Suprasegmental Features," in *ICSLP'98*, Sydney, Australia, 1998, vol. 4, p. 1403.

[20] H. Wright, M. Poesio, and S. Isard, "Using High Level Dialogue Information for Dialogue Act Recognition using Prosodic Features," in *ESCA Workshop on Prosody and Dialogue*, Eindhoven, Holland, September 1999.

[21] S. Keizer, Akker. R., and A. Nijholt, "Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues," in *3rd ACL/SIGdial Workshop on Discourse and Dialogue*, Philadelphia, USA, July 2002, pp. 88–94.

[22] G. Ji and J. Bilmes, "Dialog Act Tagging Using Graphical Models," in *ICASSP'05*, Philadelphia, USA, March 2005, vol. 1, pp. 33–36.

[23] N. Reithinger and E. Maier, "Utilizing Statistical Dialogue Act Processing in VERBMOBIL," in *33rd annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 1995, pp. 116–121, Association for Computational Linguistics.

[24] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, pages 500-528, MIT Press. Cambridge, MA, May 1999.

[25] V. Strom, "Detection of Accents, Phrase Boundaries and Sentence Modality in German with Prosodic Features," in *Eurospeech'95*, Madrid, Spain, 1995.

[26] P. Král, C. Cerisara, and J. Klečková, "Combination of Classifiers for Automatic Recognition of Dialog Acts," in *Interspeech'2005*, Lisboa, Portugal, September 2005, pp. 825–828, ISCA.