Sentence Structure for Dialog Act Recognition in Czech

Pavel Král, Christophe Cerisara
LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
France
e-mail: {kral,cerisara}@loria.fr

Jana Klečková, Tomáš Pavelka
Dept. Informatics & Computer Science
University of West Bohemia
Plzeň, Czech Republic
e-mail: {kleckova,tpavelka}@kiv.zcu.cz

Abstract

This paper deals with automatic dialog acts (DAs) recognition in Czech based on sentence structure. We consider the following DAs: statements, orders, yes/no questions and other questions. In our previous works, we have proposed, implemented and evaluated new approaches to automatic DAs recognition based on sentence structure and prosody. The word sequences were manually transcribed. The main goal of this paper is to evaluate the performances of our approaches when these word sequences are unknown and estimated from a speech recognizer. Our system is tested on a Czech corpus that simulates a task of train tickets reservation. When manual transcription is used, classification accuracy without and with sentence structure models is 91 %, 94 % and 95 %. The recognition accuracy reaches 96 % with prosodic combination. When word sequences are estimated from a speech recognizer, the classification score is 88 % without and 91 % and 92 % with sentence structure models. The combination with prosody gives 93 % of accuracy.

1. Introduction

A *dialog act (DA)* represents the meaning of an utterance at the level of illocutionary force [1].

For example, "question" and "answer" are both possible dialog acts. Automatically recognizing such dialog acts is of crucial importance to interpret and guarantee natural user interactions.

The main goal of this paper is to compare the classification accuracy of automatic DA recognition approaches, when manual word transcription is used and when word sequences are estimated from a speech recognizer.

The dialog acts recognition module is designed to be integrated into a dialog system. Such a system shall exploit dialog acts to better interpret the user's inputs. Our main interest is question detection, because it is an important clue for dialogue management. For example, when our system detects an explicit question, it has to treat it immediately and react accordingly.

Section 2 presents a short review of dialog acts recognition approaches. Section 3 presents our methods based on sentence structure. Section 4 describes the LASER speech recognizer [2]. Section 5 evaluates and compares these methods in two different cases: when manual word transcription is used, and when the word transcription is estimated from a speech recognizer. In the last section, we discuss the research results and we propose some future research directions.

2. Short review of dialog acts recognition approaches

To the best of our knowledge, there is very little existing work on automatic modeling and recognition of dialog acts in the Czech language. Alternatively, a number of studies have been published for other languages, and particularly for English and German.

In most of these works, the first step consists to define the set of dialog acts to recognize. In [3], [4], 42 dialog acts classes are defined for English, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [5]. Jekat [6] defines for German and for Japanese in VERBMOBIL 42 DAs, with 18 DAs at the illocutionary level. The MALTUS (Multidimensional Abstract Layered Tagset for Utterances) [7] is another DAs tag set based on DAMSL.

Automatic recognition of dialog acts is usually realized using one of, or a combination of the three following models:

1. DA-specific language models
2. dialog grammar
3. DA-specific prosodic models

The first class of models infers the DA from the word sequence. Usually, probabilistic approaches are represented by language models such as n-gram [4], [8], or knowledge based approaches such as semantic classification trees [8].

The methods based on probabilistic language models exploit the fact that different DAs use distinctive words. Some cue words and phrases can serve as explicit indicators of dialogue structure. For example, 88.4 % of the trigrams "<start> do you" occur in English in *yes/no questions* [9].

Semantic classification trees are decision trees that operate on word sequence with rule-based decision. These rules are trained automatically on a corpus. Alternatively, in classical rule based systems, these rules can be coded manually.

A dialog grammar is used to predict the most probable next dialog act based on the previous ones. It can be modeled by hidden Markov models

(HMMs) [4], Bayesian Networks [10], Discriminative Dynamic Bayesian Networks (DBNs) [11], or n-gram language models [12].

Prosodic models [3] can be used to provide additional clues to classify sentences in terms of DAs. A lexical and prosodic classifiers are combined in [4].

3. Dialog act recognition approaches

Syntax information is often modeled by probabilistic n-gram models. However, these n-grams usually model *local* sentence structure only.

In our system we propose to include information related to the position of the words within the sentence. This method presents the advantage of introducing valuable information related to the *global* sentence structure, without increasing the complexity of the overall system.

3.1. Sentence structure model

The general problem of automatic DAs recognition is to compute the probability that a sentence belongs to a given dialog act class, given the lexical and syntactic information, i.e. the words sequence.

We simplify this problem by assuming that each word is independent on the other words, but is dependent on its position in the sentence, 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 dialog act class of the test sentence, w represents a word and P its position in the sentence.

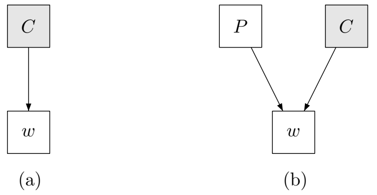

Figure 1. Graphical model of our approaches: grayed nodes are hidden

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 sentence is given by equation 1.

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

This system is referred to as “unigram”.

On the right part of figure 1, information about the position of each word is included. However, this model poses two practical issues that have to be solved:

- Sentences have different length.
- New variable P greatly reduces the ratio between the size of the corpus and the number of free parameters to train.

We propose in [13] two methods to solve these problems. The first one, *multiscale position* method, exploits a description of the sentence in several levels to smooth the probabilities across these levels. The second one, *non-linear merging* method, models the dependency between W and P by a non-linear function that includes P.

3.2. Combination with prosody

Following the conclusions of previous studies [14], only the two most important prosodic attributes are used: F0 and energy. Let us call F the set of prosodic features for one sentence. We use a Gaussian Mixture Model (GMM) classifier that computes $P(F|C)$.

The outputs of the lexical, position and prosodic models are normalized to the interval $[0;1]$. They respectively approximate $P(C|W)$, $P(C|W,P)$ and $P(C|F)$.

These probabilities are then combined with a Multi-Layer Perceptron (MLP), as suggested in our previous works [15].

4. LASER speech recognizer

The LASER (LICS Automatic Speech Extraction/Recognition) software is currently under development by the Laboratory of Intelligent Communication Systems (LICS), University of West Bohemia. The goal is to develop a set of tools that would allow training of acoustic models and recognition with task dependent grammars or more general language models.

The architecture is based on a so called *hybrid* framework which combines the advantages of hidden Markov model approach with the universality of artificial neural networks. A typical hybrid system uses HMMs with state emission probabilities computed from output neuron activations of a neural network (such as the multi layer perceptron).

4.1. Neural network acoustic model

According to many authors (see e.g [16]) the use of a neural network for the task of acoustic modeling has several potential advantages over the conventional Gaussian mixtures seen in today’s state-of-the-art recognition systems. Among the most notable ones are its economy – a neural network has been observed to require less trainable parameters to achieve the same recognition accuracy as a Gaussian mixture model and context sensitivity – the ability to include features from several subsequent speech frames and thus incorporate contextual information.

A three layer perceptron serves as an acoustic model in the latest version of the recognizer. It has 117 input neurons (there are 13 MFCC coefficients per speech frame and 9 subsequent frames are used as features), 400 hidden neurons and 36 output neurons corresponding to our choice of 36 context independent phonetic units (which roughly correspond to Czech phonemes). Experiments with larger hidden layer sizes have been carried out but the 400 hidden neurons were chosen as a good trade-off between modeling accuracy and computational requirements.

The incremental version of the back-propagation algorithm has been found as the fastest converging training strategy for this task. Also in order to further speed up the convergence cross entropy error criterion is used instead of the usual summed square error. The training of a multi layer perceptron requires the exact locations of phoneme boundaries to be known, i.e. it must be known for each speech frame in the training set to which phonetic class it belongs to. These can be obtained via forced Viterbi alignment from the transcriptions of the training utterances. An already trained recognizer is necessary for this process. It is also beneficial to generate a new set of phonetic labels using the newly trained hybrid recognizer and repeat the training process once more.

Similarly to other automatic speech recognition systems there are three-state HMMs modeling phonetic unit. However all three states share the same emission probability computed from the activation value of a neuron in the output layer of the MLP. This can be viewed as a minimum phoneme duration constraint which (according to our experiments) significantly increases recognition accuracy. Because each state is tied to a neuron representing one phonetic class the outputs of a well trained MLP can be interpreted as state posterior probabilities $P(S_j|o)$[1] which can be (by the application of Bayes' rule) changed to state emission probabilities (S_j denotes j-th HMM state):

$$P(o|S_j) = \frac{P(S_j|o) \cdot P(o)}{P(S_j)}. \tag{2}$$

The term $P(o)$ remains constant during the whole recognition process and hence can be ignored so the emission probabilities can be acquired by dividing the network outputs by the class priors (relative frequencies of each class observed in training data).

The HMM state transition probabilities are not trained since their contribution to recognition accuracy is negligible (in speech recognition applications, according to our experiments). Uniform distribution is assumed instead.

4.2. Language model

Our biggest problem is the size of the training data since the training set for the language model consists of the transcriptions of the railway corpus (see section 5.1). Experiments with pure n-gram models led to poor recognition performance. Our solution is to merge words into classes and construct an n-gram model based on those classes. This should compensate the lack of training data for infrequent n-grams.

The method tries to automatically cluster words into classes according to their functional position in sentence. The algorithm (see [17]) starts with assigning each words into separate class and then starts merging two classes at a time. The process is stopped when there is a desired number of classes.

The larger the number of word classes used in language model training the more it reflects the syntactical structure of the training set and, since the training set is small, hinders the recognition of n-grams not seen in the training data. A compromise leading to best performance on the test set was to cluster the 1400 words in the railway corpus into 100 classes and use those to train a trigram language model.

4.3. Decoding

The acoustic model provides a local match, i.e. it estimates a score for each phonetic unit in a short speech frame. The percentage of correctly recognized frames (those where the correct phonetic unit gets the highest score) at this point is around 70 %. The role of the decoder is to search those scores and output the most likely word sequence. The usual technique for this is the Viterbi algorithm (see e.g. [18]).

Our initial experiments were carried out with hand written grammars describing all possible recognized utterances. The advantage of this approach is that the search space is relatively small and it is not necessary to do any pruning during the search.

When long span language models (such as trigrams) together with larger dictionaries are used it is no longer possible to do an exhaustive search. In order to lower the computational complexity the number of active states (i.e. those that will take part in further computation) must be pruned. A simple but efficient pruning method is the *beam search* which prunes all states having lower score than a given percentage of the highest score. Such search is no longer *admissive*, i.e. does not guarantee to find the most likely word sequence. Despite this it has been shown to work well in practice.

Another way to speed up the search is to reorganize the dictionary. For example if two words start with the same phoneme (phonetic unit) than the computation for their initial states needs to be done only once. This leads to a tree like structure of the dictionary and its respective HMM. While the tree reduces the size of the HMM to about 40 % of its original size, it can reduce the number of active nodes during beam search by an order of magnitude.

5. Experiments

5.1. Dialog acts corpus

Czech Railways corpus, which contains human-human dialogs, is used to validate the proposed methods. The number of sentences of this corpus is shown

[1] In HMM terminology o represents observation, i.e in our case the feature vector

in column 2 of table 1.

The LASER recognizer is trained on 6234 sentences (c.f. first part of table 1), while 2173 sentences pronounced by different speakers (c.f. second part of table 1) is used for testing. Sentences in the testing part of the corpus has been labelled manually with the following dialog acts: statements (S), orders (O), yes/no questions (Q[y/n]) and other questions (Q). The word transcription estimated from the LASER recognizer is used to compare the performances of DAs recognition experiments with the scores obtained by manual word transcription.

All experiments for DAs recognition 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 ± 1 %.

DA	No.	Example	English translation
1. Training part			
Sent.	6234		
2. Testing part (labeled by DAs)			
S	566	Chtěl bych jet do Písku.	I would like to go to Písek.
O	125	Najdi další vlak do Plzně!	Look at for the next train to Plzeň!
Q[y/n	282	Řekl byste nám další spojení?	Do you say next connection?
Q	1200	Jak se dostanu do Šumperka?	How can I go to Šumperk?
Sent.	2173		

Table 1. Composition of Czech Railways corpus

5.2. Sentence structure experiments

The first part of table 2 shows the recognition score obtained with a unigram model.

The recognition accuracy of the sentence structure models are shown in the second part of table 2. The global recognition accuracy of *multiscale position* model is 93.8 % and of *non-linear merging* model is 94.7 %, which is the best score obtained by every module taken individually.

Non-linear merging model is implemented by a Neural Network of type Multi-Layer Perceptron (MLP). The chosen MLP topology is composed from three layers: 4 (for each DA class) times 8 (equal-size segments of the sentence) inputs, 12 neurons in hidden layer and 4 output neurons, which encode the *a posteriori* class probability.

5.3. Combination with prosody

The third section of table 2 shows the recognition score of the prosodic GMM. The best recognition accuracy is obtained with a 3-mixtures GMM.

The last part of table 2 shows the recognition results when the prosodic GMM and the MLP-position models (described in [15]) are combined with another MLP.

The combination of models gives better recognition accuracy 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	Q[y/n]	Q	Global
1. Lexical information					
Unigram	93.5	77.6	96.5	89.9	**91.0**
2. Sentence structure					
Multiscale	94.7	70.4	96.1	95.3	**93.8**
Non-linear	90.3	83.2	91.1	98.8	**94.7**
3. Prosodic information					
GMM	47.7	43.2	40.8	44.3	**44.7**
4. Combination					
MLP	91.5	85.6	94.0	98.7	**95.7**

Table 2. Dialog acts recognition accuracy for different approaches/classifiers and their combination with manual word transcription

5.4. Recognition with LASER recognizer

Table 3 shows DAs recognition scores, when word transcription is estimated by the LASER recognizer. The results are obtained with word class based trigram language model (see section 4.2). Sentence recognition accuracy is 39.78 % and word recognition accuracy is 83.36 %.

Table 3 structure is the same as table 2.

6. Conclusions

In this work, we compared the performances of several methods for automatic DAs recognition in two cases: when manual word transcription is used, and when word sequences are estimated from the LASER speech recognizer. The objective of this work was to integrate these methods into a multi-modal ticketing reservation system.

We show that the DA recognition accuracy only slightly decreases, when word sequences are estimated automatically from the recognizer. The absolute decrease of the recognition score is about 3 % only, which is insignificant for our application.

The main perspective of our work is to add dialog

	accuracy in [%]				
Approach/ Classifier	S	O	Q[y/n]	Q	Global
1. Lexical information					
Unigram	93.1	68.8	94.7	86.3	**88.2**
2. Sentence structure					
Multiscale	93.8	63.2	92.9	92.9	**91.4**
Non-linear	85.5	72.0	86.8	98.0	**91.8**
3. Prosodic information					
GMM	47.7	43.2	40.8	44.3	**44.7**
4. Combination					
MLP	88.5	77.6	90.4	97.3	**93.0**

Table 3. Dialog acts recognition accuracy for different approaches/classifiers and their combination with word transcription from LASER recognizer

history (c.f. section 2) to improve DAs recognition accuracy.

Finally, in real applications, other clues such as the current dialog state shall also be considered. However, we proposed in this work a DA recognition module that is independent from the task, and which can be easily retrained on another corpus. Another perspective is also to test these methods on another corpus (radio), another language (French) and with more DA classes.

7. References

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

[2] K. Ekštein and T. Pavelka, "Lingvo/laser: Prototyping concept of dialogue information system with spreading knowledge," in *NLUCS'04*, Porto, Portugal, April 2004, pp. 159–168.

[3] 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.

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

[5] J. Allen and M. Core, "Draft of Damsl: Dialog Act Markup in Several Layers," 1997.

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

[7] A. Clark and A. Popescu-Belis, "Multi-level Dialogue Act Tags," in *5th SIGdial Workshop on Discourse and Dialogue*, Boston MA, 2004.

[8] 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.

[9] 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.

[10] 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, PA, July 2002, pp. 88–94.

[11] G. Ji and J. Bilmes, "Dialog Act Tagging Using Graphical Models," in *ICASSP'05*, Philadelphia, March 2005.

[12] 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.

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

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

[15] 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.

[16] H. Bourlard and N. Morgan, "Hybrid hmm/ann systems for speech recognition: Overview and new research directions," in *Summer School on Neural Networks*, 1997, pp. 389–417.

[17] J.F. Allen, *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, 1988.

[18] S. Young and al., *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering, 2002.