AUTOMATIC DIALOG ACTS RECOGNITION BASED ON SENTENCE STRUCTURE

Pavel Král, Christophe Cerisara

LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
France
{kral,cerisara}@loria.fr

Jana Klečková

Dept. Informatics & Computer Science
University of West Bohemia
Plzeň, Czech Republic
kleckova@kiv.zcu.cz

ABSTRACT

This paper deals with automatic dialog acts (DAs) recognition in Czech. Our work focuses on two applications: a multimodal reservation system and an animated talking head for hearing-impaired people. In that context, we consider the following DAs: statements, orders, investigation questions and other questions. The main goal of this paper is to propose, implement and evaluate new approaches to automatic DAs recognition based on sentence structure and prosody. Our system is tested on a Czech corpus that simulates a task of train tickets reservation. With lexical-only information, the classification accuracy is 91 %. We proposed two methods to include sentence structure information, which respectively give 94 % and 95 %. When prosodic information is further considered, the recognition accuracy reaches 96 %.

1. INTRODUCTION

Dialog acts (DAs) are defined by Austin [1] as a meaning of an utterance at the level of illocutionary force. There are many possible different dialog acts, but for our applications, we are mainly interested in recognizing questions. Such an information can be used to help dialog systems to identify explicit user requests and implicit interactions. It may also be useful to animate a talking head [2] that reproduces someone's speech. For example, a question mark near the talking head is displayed or the brows are raised when a question is asked. This improves the naturalness of the talking head, and further convey additional para-linguistic informations that are not present otherwise.

We proposed and compared in [3] several methods to combine prosodic and lexical classifiers for DA recognition. We now extend this preliminary work by proposing two novel approaches to further include sentence structure information in our automatic DA recognizer.

Section 2 presents a short review of dialog acts recognition approaches, with a focus on syntax and sentence structure information. Section 3 describes two new methods to take into account sentence structure. Section 4 evaluates and compares these methods. 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 are very few 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 [4, 5], 42 dialog acts classes are defined for English, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [6]. Jekat [7] 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) [8] 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 words sequence. Usually, probabilistic approaches are represented by language models such as n-gram [5, 9], or knowledge based approaches such as semantic classification trees [9].

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 *investigation questions* [10].

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) [5], Bayesian Networks [11], Discriminative Dynamic Bayesian Networks (DBNs) [12], or n-gram language models [13].

Prosodic models [4] can be used to provide additional clues to classify sentences in terms of DAs. A lexical and prosodic classifiers are combined in [5].

Another classes of approaches use *multi-level* information to automatically recognize of DAs. Rosset [14] assumes that the word position is more important than the exact word identity. Therefore the first word only is used as lexical information. The following remaining multi-level information are computed: speaker identification, DAs history and number of utterance units in each turn.

Unsupervised DA recognition approaches have also been proposed [15]. They are based on Kohonen Self-Organizing Feature Maps and superficial utterance features, such as speaker identity, sentence mood or type of subject of an utterance.

3. DIALOG ACT RECOGNITION FROM PROSODY, LEXICON AND SENTENCE STRUCTURE

When considered, syntax information is often modeled by probabilistic n-gram models in automatic DA recognition systems. However, these n-grams usually model local structures only. Syntax parsing could be used to associate sentence structures to particular dialog acts, but conceiving general grammars is still an open issue, especially for spontaneous speech.

We propose to include in our system a simplified information related to the structure of sentences, i.e. 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 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 from the other words, but is dependent from 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. On this figure, C encodes the dialog act class of the test sentence, w represents a word and P its position in the sentence.

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:

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

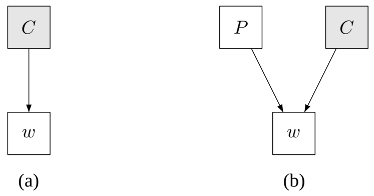

Fig. 1. Graphical model of our approaches: grayed nodes are hidden

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.

First, sentences have different length. A constant number of different positions N_P is fixed, and N_P probabilities are computed for each sentence. Let us call T the actual number of words in the sentence. The T words are aligned linearly with the N_P positions. Two cases may occur:

- When $T \leq N_P$, the same word may appear at several positions.
- When $T > N_P$, several words can be aligned with one position. These words are replaced by a single “hyperword” h whose probability is the average over the N_i words $(w_i)_{N_i}$ involved:

$$P(h|C,P) = \frac{1}{N_i} \sum_{i}^{N_i} P(w_i|C,P) \qquad (2)$$

Second, this new variable P greatly reduces the ratio between the size of the corpus and the number of free parameters to train. We propose two methods to solve this issue: the first one exploits a multiscale description of the sentence to smooth the probabilities across the scales, while the second one models the dependency between W and P by a non-linear function that includes P.

3.1.1. Multiscale position

In this approach, P can take a different number of values depending on the scale. All these scales can be represented on a tree, as shown in figure 2. At the root of the tree (coarse scale), P can take only one value: it is equivalent to unigrams. Then, recursively, sentences are split into two parts of equal size and the number of possible positions is doubled.

For each word w_i, a threshold is applied on its number of occurences and $P(w_i|C,P)$ for this word is computed at the finest scale that contains a greater number of occurences. This corresponds to the standard back-off technique [16] to solve the problem of lack of data.

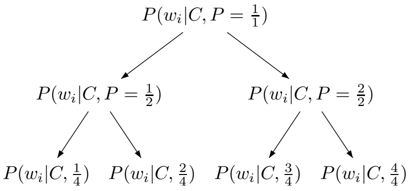

Fig. 2. Multiscale position tree

3.1.2. Non-linear merging

In this approach, unigram probabilities are computed for each word and passed to a muti-layer perceptron (MLP), where the position of each word is encoded by its input index: the i^{th} word in the sentence is filled into the i^{th} input of the MLP. The output of the MLP corresponds to the *a posteriori* probabilities $P(C|W)$.

3.2. Prosody

Following the conclusions of previous studies [17], 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 test two classifiers: a MLP that computes $P(C|F)$ and a Gaussian mixture model (GMM) that models $P(F|C)$.

3.3. Combination

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

These probabilities are then combined with another MLP, as suggested in our previous experiments [3].

4. EXPERIMENTS

4.1. Dialog acts corpus

A subset of the Czech Railways corpus, which contains human-human dialogs, is used to validate the proposed methods. For the next experiments, it has been labelled manually with the following set of dialog acts: statements, orders, investigation questions and other questions. The corpus contains 2173 utterances (566 statements (S), 125 orders (O), 282 investigation questions (Q[y/n]) and 1200 others questions (Q)). All the following 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 $< 1\%$.

4.2. Sentence structure experiments

Figure 3 shows the recognition accuracy of the sentence structure model in function of the minimum number of word occurences, which defines the threshold used in the multiscale tree. The depth of the tree used in this experiment is 3, which defines 8 segments. The unigram model recognition accuracy is also reported on this figure.

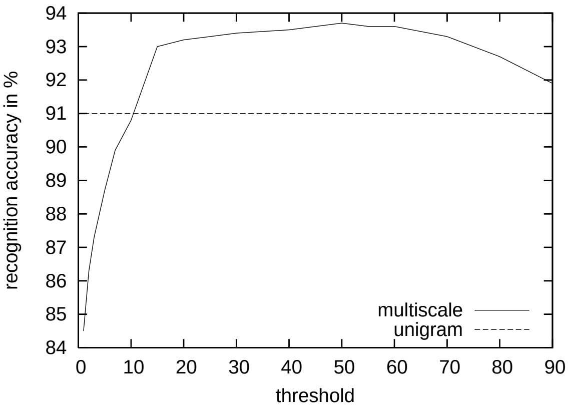

Fig. 3. Dialog acts recognition accuracy of the multiscale position tree system. The X-axis represents the minimum number of words in the tree, and the Y-axis plots the DA recognition accuracy

The second proposed model that takes into account position information uses a MLP to merge the unigram probabilities and their position. As with the tree smoothing approach, each sentence is split into 8 equal-size segments. The MLP has thus 4 (for each DA class) times 8 inputs. 12 neurons populate the intermediate layer, and 4 output neurons encode the *a posteriori* class probability.

The global recognition accuracy of this model is 94.7 %. This is the best result obtained by every module taken individually.

4.3. Prosody

Table 1 compares the recognition accuracy of the prosodic GMM and MLP. The best MLP topology uses three layers: 40 inputs, 18 neurons in hidden layer and 4 outputs. The best recognition accuracy is obtained with a 3-mixtures GMM.

These recognition scores are much lower than the ones obtained with sentence structure, but our objective is to show that prosody may nevertheless bring some relevant clues that are not related to words sequence.

4.4. Combination

The last part of table 1 shows the recognition results when the prosodic GMM and the MLP-position models (described in 3.1.2) are combined with another MLP.

The combination of 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	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**
MLP	38.7	49.6	52.6	34.0	**43.5**
4. Combination					
MLP	91.5	85.6	94.0	98.7	**95.7**

Table 1. Dialog acts recognition accuracy for different approaches/classifiers and their combination in %

5. CONCLUSIONS

In this work, we presented two new methods for automatic DAs recognition, with the objective to integrate them into two target applications: a multimodal ticketing reservation system, and an animated talking head. We show that the DA recognition accuracy increases when sentence structure information is used, compared to lexical models. We further compare two approaches to model words position in sentences.

The first perspective of this work consists to use an automatic speech recognizer instead of the manual word transcriptions used in our experiments. The errors coming from the speech recognizer may temper the dominant position of the lexical and sentence structure classifiers. 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.

6. REFERENCES

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

[2] P. Král and J. Klečková, "Speech Recognition and Animation of Talking Head," in *IWSSIP'03*, Prague, Czech Republic, September 2003.

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

[4] 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.

[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," 1997.

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

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] S. Rosset and L. Lamel, "Automatic Detection of Dialog Acts Based on Multi-level Information," in *Interspeech'2004 - ICSLP*, Jeju Island, October 2004, pp. 540–543.

[15] T. Andernach, M. Poel, and E. Salomons, "Finding Classes of Dialogue Utterances with Kohonen Networks," in *ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, Prague, Czech Republic, April 1997, pp. 85–94.

[16] J. Bilmes and K. Kirchhoff, "Factored Language Models and Generalized Parallel Backoff," in *Human Language Technology Conference*, Edmonton, Canada, 2003.

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