Combination of classifiers for automatic recognition of dialog acts

Pavel Král[1], *Christophe Cerisara*[1], *Jana Klečková*[2]

[1]LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
FRANCE
[2]Dept. Informatics & Computer Science
University of West Bohemia
Plzeň, Czech Republic
kral,cerisara@loria.fr,kleckova@kiv.zcu.cz

Abstract

This paper deals with automatic dialog acts (DAs) recognition in Czech. The dialog acts are sentence-level labels that represent different states of a dialogue, depending on the application. 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, yes/no questions and other questions. We propose to use both lexical and prosodic information for DAs recognition. The main goal of this paper is to compare different methods to combine the results of both classifiers. On a Czech corpus simulating a reservation of train tickets, the lexical information only gives about 92 % of classification accuracy, while prosody gives only about 45 % of accuracy. When both classifiers are combined with a multilayer perceptron, the lowest (lexical) word error rate further decreases by 26 %. We show that this improvement is close to the optimal one, given the correlation of the lexical and prosodic features. The other combination schemes do not outperform the lexical-only results.

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 goal of this paper is to recognize dialog acts by combining two sources of information: lexical and prosodic. Our main contribution concerns the comparison of different combination methods.

In this work, the dialog acts recognition module is designed to be integrated into the two following applications. The first one is a dialog system that handles reservation tasks, and the second one deals with the animation of a talking head. The dialog system shall exploit dialog acts to better interpret the user's inputs, while the talking head, which reproduces visually someone's speech for hearing-impaired people, shall benefit from the dialog act labels to animate the 3D face more naturally.

The following section presents an overview of some existing works in this domain. Next, our approach to dialog acts recognition is explained. The first classifier uses lexical information, the second one prosody and the last one focuses on the combination of both previous techniques. In section 4, we evaluate and compare 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 [2, 3, 4], 42 dialog acts classes are defined, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [5]. This list is usually reduced into a much smaller number of broad classes, because some classes occur only seldom, and because all these classes are not needed for dialog understanding. A common regrouping is the following [2]:

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

Automatic recognition of these dialog acts can then be achieved using one of, or a combination of the three following models:

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

The first class of models infer the DA associated to a sentence from its words sequence. They generally use probabilistic language models such as n-gram [3, 6], semantic classification trees [6], or neural networks [7, 8, 4]. This lexical information usually contributes the most to characterize the sentence DA.

Prosodic models are often used to provide additional clues to classify sentences in terms of DAs. The dialog acts can be characterized by prosody as follows [9]:

- a falling intonation for a statement
- a rising F0 contour for a question (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

Prosody is used for DAs recognition in [2, 3, 4, 10, 11, 12]. In [2], the duration, pause, fundamental frequency (F0), energy and speaking rate prosodic attributes are modeled by a CART-style decision trees classifier. In [10], prosody is used to segment utterance. The duration, pause, F0-contour and energy features are used in [11, 12]. These two studies compute several features 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.

Very often, a dialog grammar is further used to predict the most probable next dialog act based on the previous ones. It can be modeled by hidden Markov models [3, 4] or Discriminative Dynamic Bayesian Networks (DBNs) [13].

The lexical and prosodic classifiers are combined in [2, 3, 4]. The following equation is used:

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

where C represents a dialog act and W and F represents respectively the lexical and prosodic information. W and F are assumed independent.

3. Approaches

Following the conclusions of the previous studies, which suggest that prosody brings some valuable information that can not be captured by the lexical models alone, this work combines lexical and prosodic classifiers to recognize DAs. The main contribution of this work concerns the use and comparison of different kinds of combination methods for this task.

3.1. Lexical information

Let us call $W = (w_1, \cdots, w_T)$ the sequence of words in a test sentence. We train a unigram classifier to model the likelihood that this words sequence belongs to a given dialog act C:

$$P(W|C) = \prod_{i=1}^{T} P(w_i|C) \quad (2)$$

3.2. Prosody

Following the conclusions of previous studies [14, 15], only the two most important prosodic attributes are considered: F0 and energy. The F0 curve is computed with the autocorrelation function. The F0 and energy values are computed on every overlapping speech window. The F0 curve is completed by linear interpolation on the unvoiced parts of the signal. Then, each sentence is decomposed into 20 segments and the average values of F0 and energy are computed within each segment. This number is chosen experimentally [15]. We thus obtain 20 values of F0 and 20 values of energy per sentence. Let us call F the set of prosodic features for one sentence.

We test two classifiers: a muti-layer perceptron (MLP) that computes $P(C|F)$ and a Gaussian mixture model (GMM) that models $P(F|C)$. Both classifiers error rates are reported in the following experiments, but as they give comparable results, the combination with the lexical classifier uses only the prosodic GMM.

3.3. Combination

The outputs of our classifiers are $P(W|C)$ for the lexical model and $P(F|C)$ for the prosodic one, where C is the dialog act class, W is the words sequence of the utterance and F represents the prosodic features of the utterance.

We first normalize these likelihoods to compute the a posteriori class probabilities:

$$P(C=c|W) = \frac{P(W|C=c).P(C=c)}{\sum_{i=1}^{N} P(W|C=i).P(C=i)} \quad (3)$$

where c and i represent DAs classes, N is the number of DAs and $P(C)$ is the *prior* probability of class C. We assume that all classes are equi-probable. A similar equation is applied to the prosodic model.

Next, several combination methods are tested. The first three ones, *maximum*, *minimum* and *median*, are based on order statistics [16]: For each class, the *a posteriori* probabilities returned by both classifiers are ordered, and the final score of each class is respectively the greatest, smallest, and average a posteriori probability for that class.

The fourth combination (*product*) assumes that both classifiers are independent and simply computes the product of their posterior probabilities:

$$P(C|W,F) \simeq P(C|W).P(C|F) \quad (4)$$

The fifth combination, *weighted linear* computes a weighted linear combination of the a posteriori probabilities:

$$P(C|W,F) \simeq (g).P(C|W) + (1-g).P(C|F) \quad (5)$$

The weight g is optimized via a grid-search on a development corpus.

The last algorithm combines the a posteriori probabilities with a MLP.

Note that the first four combinations are "unsupervised" while the last two ones are "supervised" and require a development corpus.

4. Experimental setup

4.1. Dialog acts corpus

A subset of the Czech Railways corpus, which contains some human-human dialogs, is used to validate the proposed methods. It was created at the University of West Bohemia mainly by members of the Department of Computer Science and Engineering. For the next experiments, it has been labelled manually with the following set of dialog acts: statements, orders, yes/no questions and other questions. This list is derived from the seven classes considered in section 2, which have been further simplified with regard to the specifics of this corpus. The corpus contains 2173 utterances (566 statements (S), 125 orders (O), 282 yes/no 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. Lexical approach

The first part of table 2 shows the recognition accuracy of the lexical classifier (a unigram model) only. The words sequence is given by manual transcription of the utterances. The global accuracy of this experiment is 92.3 %. This score confirms that the most important information to recognize DAs is given by the words sequence.

4.3. Prosodic approach

The middle part of table 2 shows the recognition accuracy with the prosodic GMM and MLP. The best recognition accuracy is obtained with the 3-mixtures GMM. It is difficult to use more Gaussians, because of the lack of training data, mainly for class O. The best MLP topology uses three layers: 40 inputs, 18 neurons in hidden layer and 4 outputs. The global accuracies of the GMM and MLP classifiers are comparable. These recognition scores are much lower than the one obtained with the lexical features, but our objective is to show that prosody may nevertheless bring some relevant clues that are not related to the words sequence.

4.4. Classifier combination

We first study the correlation matrix of both lexical and prosodic (GMM only) classifiers in table 1: this matrix shows the ratio of the examples that are simultaneously correctly or incorrectly classified by both classifiers. For example, 40.04 % of the examples are classified correctly by both classifiers while 5.57 % of the examples are not recognized by any classifier. An interesting remark from this table is that 2.12 % of the examples are recognized by the prosodic classifier, but not by the lexical one. This suggests that there is a small but significant potential improvement that can be obtained by considering prosodic information as well.

	lexical correct	lexical incorrect
prosodic correct	40.04	2.12
prosodic incorrect	52.28	5.57

Table 1: Correlation of classification error rate of both classifiers in %

Let us now study the last part of table 2 that shows the recognition results when combining both classifiers (the prosodic GMM model is used).

We can note that, amongst order statistics combiners, the minimum and median ones are better than the maximum one. But we can also observe that every unsupervised combination gives a lower accuracy than the lexical classifier alone. This can be explained by the fact that we combine only two classifiers, and most importantly because of the big difference between each individual classifier recognition accuracy. Indeed, this is confirmed by the *weighted linear* combination, which optimal weight is 0.97 in favor of the lexical approach.

The best recognition accuracy is obtained with the MLP combination, which reduces the lexical word error rate by an absolute 2 %. This figure can be compared with the 2.12 % shown in table 1.

	ACC in [%]				
Approach/ Classifier	S	O	Q[y/n]	Q	Global
Lexical information					
Unigram	88.5	90.4	92.9	94.2	92.3
Prosodic information					
GMM	47.7	43.2	40.8	44.3	44.7
MLP	38.7	49.6	52.6	34.0	43.5
Combination of approaches					
Maximum	81.8	81.6	88.3	57.9	69.4
Minimum	80.0	73.6	84.8	64.6	71.7
Median	81.3	81.6	88.3	63.2	72.2
Product	81.1	76.8	86.2	64.4	72.3
Weighted Linear	88.5	90.4	92.9	94.2	92.3
MLP	90.3	88.0	92.9	97.3	94.3

Table 2: Dialog acts ACC for different approaches/classifiers and theirs combinations in %

5. Conclusions

In this work, we have studied and compared different methods to combine lexical and prosodic information in the context of automatic dialog act recognition, with the objective to integrate this approach into two applications: a multimodal ticketing reservation system, and an animated talking head.

The lexical knowledge source is the most important one: on a Czech corpus that simulates the first application it already recognizes correctly about 92 % of the DAs. However, we showed that it is possible to improve this baseline result by combining this lexical classifier with a prosodic one with a MLP. Then, a statistically significant 2 % absolute improvement can be achieved, which is actually very close to the potential improvement derived from the correlation matrix between both classifiers. This confirms that prosodic clues are *complementary* to the lexical ones, as it has been already suggested in other studies such as [2].

All the other combination schemes, and in particular the unsupervised ones, do not reach the level of the lexical classifier alone. This shows the importance to finetune the combiner on a development corpus in our experimental set-up. This might result from the large difference in the performances of both classifiers, and also from the small number of experts that are combined.

The first perspective of this work will consist to use an automatic speech recognizer, such as the one described in [17], 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 classifier. Another interesting development shall be to combine more classifiers, which may favor the other combination schemes than the MLP. 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] 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.

[3] A. Stolcke *et al.*, "Dialog act modeling for conversational speech," in *AAAI Spring Symp. on Appl. Machine Learning to Discourse Processing*, 1998, pp. 98–105.

[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] 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.

[7] H.-F. Wang, W. Gao, and S. Li, "Dialog acts analysis of spoken chinese based on neural networks," *Chinese Journal of Computers*, 1999.

[8] 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.

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

[10] 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, 1996.

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

[12] 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.

[13] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *ICASSP'05*, Philadelphia, March 2005.

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

[15] J. Kleckova and V. Matousek, "Using Prosodic Characteristics in Czech Dialog System," in *Interact'97*, 1997.

[16] K. Tumer and J. Ghost, "Robust combining of disparate classifiers through order statistics," *Computer Science*, May 1999.

[17] 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.