University of West Bohemia
Faculty of Applied Sciences

Lexical Information, Syntax and Semantics for Natural Language Processing

Habilitation Thesis

Pilsen, December 2014

Pavel Král

Acknowledgements

I wish to express my thanks to my colleagues who have participated in the work related with this thesis particularly to the co-authors of the papers mentioned below and to the those who were involved in the design, implementation and testing of the supporting applications.

I would like also thank to my wife Dana and to my daughters Anežka and Veronika for their support, love, understanding and patience during this work.

Finally, I wish to thank my friends Ladislav Lenc, Miloslav Konopík, Aurelie Deveau and Emmanuel Belut who have participated in the English corrections.

Abstract

Natural language processing is an interdisciplinary research field dealing with human-computer interaction in natural language. It involves many scientific challenges as for instance natural language understanding, document classification, machine translation, punctuation generation, dialogue management, automatic dialogue act recognition and many others.

This habilitation thesis shows the importance of lexical information, syntax and semantics to improve the results of the three above mentioned important natural language processing tasks: dialogue act recognition, punctuation generation and automatic document classification. Several novel approaches in these research areas are proposed, implemented and experimentally validated on the different data-sets mainly in the Czech language.

We demonstrate that lexical information plays a crucial role in all these jobs. We further show that syntax is very important for dialogue act recognition and for punctuation generation. We also prove that semantic information is beneficial to improve the document classification accuracy.

Another important scientific contribution consists in proposing a novel semi-supervised parsing approach based on a discriminative graphical model. The proposed method uses only a few manual rules as constraints instead of a few labelled sentences.

The author's research work has also an important impact on the commercial usage as proved, for instance, by the experimental version of the document classification system which is already used by the Czech News Agency.

This thesis is a collection of seven articles. These papers are first shortly presented with their main contributions in the textual part of this document.

Index terms— Czech News Agency; Dialogue Act; Document Classification; Language Model; Lexical Information; Natural Language Processing; Parsing; Punctuation Generation; Semantics; Sentence Structure; Syntax

Contents

1 Introduction

Natural Language Processing (NLP) [1] is an interdisciplinary research field dealing with human-computer interaction in natural language. This area connects three main scientific disciplines: artificial intelligence, linguistics and computational linguistics. NLP involves many research challenges such as natural language understanding (inferring the discourse meaning from an input in natural language), machine translation (automatic text translation from one language to another one), dialogue management (understanding the user's intention in the dialogue and providing an appropriate response) and many others.

This research field became very popular in the last decade, since with increasing amount and quality of linguistic data, significant progress in statistical machine learning and sufficient computation power many successful commercial NLP applications are emerging.

This thesis is focused on the usage of lexical information, syntax and semantics [2] to improve the results of three important natural language processing tasks: dialogue act recognition, punctuation generation and automatic document classification. We would like to demonstrate that lexical information plays a crucial role in all above mentioned jobs. We want further to show that syntax is more important for dialogue act recognition and for punctuation generation, rather than for document classification. However, we wish also to prove that semantic information is beneficial to improve the document classification accuracy.

1.1 About Lexis, Syntax and Semantics

In order to facilitate reading of this thesis, the three information sources that we used are briefly defined and described below.

Lexis term is originated from Greek and means the whole set of words in a language. *Lexical information* refers to the words and their positions in sentences, paragraphs or in documents. Since it represents a rich source of information, it is used in the majority of NLP tasks [3].

Syntax is a linguistic discipline dealing with the relationships between words in a sentence (i.e. clauses), the correct creation of the sentence structure and with word order. In the traditional concept, syntax is a part of the grammar together with morphology. It is not surprising that this knowledge contained in the sentence structure can also help in many speech and natural language processing applications. However, mainly due to the difficulty of obtaining this information fully-automatically, it is often used in a reduced form (word n-grams or other local structures, etc.) [4, 5].

Semantics term is also originated from the Greek language. This term comes from the word "sēmantikós" which means *significant* and is a study of the meaning. The meaning can be studied at different language levels as for instance the morphemes, words, phrases, sentences, and so on. Automatically obtaining this information is very challenging, however it can significantly increase the performance of many NLP systems [6].

1.2 Motivation and Objectives

We further shortly introduce the NLP tasks on which this document focuses. Our motivations and objectives will be also described in this section.

A *dialogue act (DA)* represents the meaning of an utterance at the level of illocutionary force [7]. In other words, the dialogue act is the function of a sentence (or its part) in a dialogue. For instance, a "question" and an "acknowledgement" are both dialogue acts.

Automatic recognition of dialogue acts is seen as a first step of dialogue understanding and is crucial to interpret users' discourse and control natural human-computer interactions. The DA labels might be used for example in a dialogue system to check whether the user is asking for some information, or to process and evaluate the user's feedback.

Automatic speech transcripts (outputs of a speech recognizer) are available often without any punctuation marks and are thus difficult to read. Moreover, a further processing as for instance parsing or information extraction is very complicated.

Therefore, automatic punctuation generation [8] is introduced in order to insert this missing sentence punctuation (full stops, commas, question marks, exclamation marks, etc.) into the "raw" texts. This task is strictly related to sentence segmentation and represents an essential feature for many speech-to-text transcription applications.

Document classification means automatic assigning of a topic (or topics) to a text document. This task becomes very important for information organization and storage because of the fast increasing amount of electronic text documents and the rapid growth of the World Wide Web.

Here we focus on the *multi-label* document classification[1] in the context of a real application for the Czech News Agency (CTK)[2]. CTK produces about one thousand text documents daily. These documents belong to different categories such as weather, politics, sport, etc. Nowadays, documents are manually annotated but this annotation is often not accurate enough. Moreover, the manual labelling represents a very time consuming and expensive task. Therefore, automatic document classification is extremely important.

The main goal of this thesis is to show the impact of lexical information, syntax and semantics in the context of three important NLP tasks. We focus on dialogue act recognition, punctuation generation and automatic

[1] One document is usually labelled with more than one label from a predefined set of labels.

[2] http://www.ctk.eu

document classification. First, we show the importance of global word position for dialogue act recognition. Three different approaches to handle this information are proposed, evaluated and compared for this purpose. The second contribution consists in proposing novel syntactic features derived from the dependency tree. These features are further integrated into the dialogue recognition system to improve resulting recognition accuracy. The next goal consists in the use of syntactic information for punctuation generation. We propose several novel syntactic features and integrate them into this task with a conditional random fields classifier. Next, we show the impact of semantic information for automatic document classification. We propose several novel features for this goal and integrate them further into a document classification system. One experimental version of this system is currently deployed in the Czech News Agency platform to simplify the current manual task. Another goal of this thesis is to propose a novel semi-supervised dependency syntactic parser based on a discriminative graphical model. This parser uses only a few manual rules as constraints instead of a few annotated sentences which are currently used.

1.3 Document Structure

Section 2 describes the syntactic parsing and our contributions into this area. The results of this task will then be used in our following work. Section 3 presents the current trends in automatic dialogue act recognition field and our contributions to this domain. Next, we shortly introduce the existing and our proposed approaches for automatic punctuation generation. Section 5 details the task of automatic document classification. We will further show out contributions into this area. In the next section, we conclude this thesis and propose the future research directions. The last section is composed of the main author's NLP papers in order to present the details of the proposed approaches and to show experimental evaluation of these methods.

2 Syntactic Parsing

Supervised dependency parsing deals with automatic construction of the annotated syntactic trees using a labelled corpus [9]. The resulting parse trees are thus highly dependent on the size and quality of the annotated corpora. A price and unavailability of such corpora for some languages have accelerated the progress of the unsupervised and semi-supervised methods.

Unsupervised parsing is a task to automatically produce syntactic trees on top of a raw, unlabelled text corpus. The most successful approaches in the field [10, 11, 12] exploit some stochastic process to decompose parents-children dependencies. The model parameters are usually trained so as to maximize the sparsity of selectional preferences in the corpus. Although no manual annotations are given, one can argue that some "linguistic" knowledge is nevertheless introduced in the model's definition, for example in the set of conditional independencies, priors and initial parameters. It has further been shown that introducing some knowledge can significantly improve the performances of unsupervised parsers and help to control their convergence and the resulting structures.

The posterior regularization framework [13] is often used to integrate constraints during inference, e.g., with a sparsity-inducing bias over unique dependency types [14] or with a few universal rules that are valid across languages [15]. Most of these works rely on a generative Bayesian model because of fundamental theoretical limitations concerning unsupervised training of discriminative models.

Semi-supervised discriminative parsers can be also very efficient. They use a few annotated sentences, such as in the SEARN paradigm [16, 17].

2.1 Contributions

In our work [18], we propose and implement a new weakly supervised parser for speech transcripts based on a discriminative directed graphical model. It exploits a few hand-crafted rules as a substitute to corpus annotations to generate constraints on the values that the latent variables can take. The parser is evaluated on English and Czech newspaper texts, and is then validated on French broadcast news transcriptions. The performances of our system are encouraging across all conditions, and match those of related state-of-the-art weakly and semi-supervised systems.

The proposed approach differs from other related works in the following aspects:

- While the majority of unsupervised parsing systems aim at producing syntactic structures on several languages without an additional effort, we rather aim at developing a framework that helps the user to build a specialized parser for his target task at a very low cost.

- An important concern in constraint-driven learning is related to the ease with which the user can design efficient constraints [19]. We argue that a natural and easy way to express constraints is by writing simple rules that describe when a specific dependency type may occur. This is precisely the intuition that led to the first rule-based systems, although in our case complex structures are obtained by combining these simple rules together and applying them in different orders. Unlike purely rule-based systems, our rules rely on a very limited definition of context and hence can be kept simple, because ambiguity is handled by the Bayesian model.

- We describe the rules creation process as an enriched annotation procedure. This better involves the rules designer within the training process, and thus leads to a joint weakly supervised training and interactive learning [20] process.

- Unsupervised parsing approaches do not produce labelled dependency arcs, because of the requirement to be generic across languages and annotation schemes. The proposed framework rather creates labelled dependency arcs for a given specific language and annotation scheme.

- Other works that exploit both hand-crafted rules and Bayesian inference combine the former on top of the latter, in the form of posterior constraints. We do the opposite, by including the rules as random variables in the inference process. This approach gives more importance to rules than other works in the field. In other words, while most previous works guide Bayesian inference with rules, we rather focus on rules and exploit a Bayesian model to help them to achieve parsing.

- Unlike many unsupervised parsers that decompose the syntactic tree with a stochastic process [10], we propose an alternative tree-decomposition paradigm, which also considers parent-children relations, but preserves the full subcategorization frame within a single random variable. Keeping the factor that involves a head and its subcategorization frame complete gives more freedom to design frame-coding adapted heuristics and, at the same time, is more rich in terms of amount of information modelled.

Selected Paper:

- C. Cerisara, A. Lorenzo, P. Král, **Weakly Supervised Parsing with Rules**, *in Interspeech 2013*, Lyon, France, August 25-29 2013, ISCA, pp. 2192-2196, ISBN: 978-1-62993-443-3.

3 Automatic Dialogue Act Recognition

Modelling and automatically identifying the structure of spontaneous dialogues are very important for better interpretation and understanding. Unfortunately, their precise representation is still an open issue, however several specific characteristics have already been clearly identified. Dialogue Acts (DAs) are one of these characteristics.

To the best of our knowledge, there is very few existing work on automatic modelling and recognition of dialogue 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 in defining a set of dialogue acts to recognize, as presented in [21, 22, 23]. Automatic recognition of dialogue acts is then usually realized using one, or a combination of the three following models:

1. DA-specific language models;
2. Dialogue grammar;
3. DA-specific prosodic models.

The first class of models infers the DA from the words sequence. Usually, probabilistic approaches are based on language models such as n-gram [21, 24], or knowledge based approaches such as semantic classification trees [24].

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* [25].

Semantic classification trees are decision trees that operate on word sequences 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 dialogue grammar is used to predict the most probable next dialogue act based on the previous ones. It can be modelled by hidden Markov models (HMMs) [21], Bayesian Networks [26], Discriminative Dynamic Bayesian Networks (DBNs) [27], or n-gram language models [28].

Prosodic models [29] can be used to provide additional clues to classify sentences in terms of DAs. A lexical and prosodic classifiers are combined in [21].

Another class of approaches use *multi-level* information to automatically recognize DAs. Rosset [30] assumes that the word position is more important than the exact word itself. Therefore, only the first word 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.

Manually annotating dialogue acts on every new corpus may be very costly and efforts have been put into developing semi-automatic methods for dialogue act tagging and discovery. Hence, the authors of [31] propose a predictive paradigm where dialogue act models are first trained on a small-size corpus and used afterwards to predict future sentences or dialogue acts. In a related vein, unsupervised dialogue act tagging of unlabelled text has recently raised a lot of attention [32, 33, 34], but we will limit ourselves in this thesis on supervised approaches.

Apart from prosodic and contextual lexical features, only a few works actually exploit syntactic relationships between words for dialogue act recognition. Some syntactic relations are captured by HMM word models, such as the widely-used n-grams [21], but these approaches only capture local syntactic relations, while we consider in our work global syntactic trees. Most other works thus focus on morphosyntactic tags, as demonstrated for instance in [35], where a smart compression technique for feature selection is introduced. The authors use a rich feature set with POS-tags included and obtain with a decision tree classifier an accuracy of 89.27%, 65.68% and 59.76% respectively on the ICSI, Switchboard and on a selection of the AMI corpus. But while POS-tags are indeed related to syntax, they do not encode actual syntactic relations.

A very few papers have nevertheless proposed some structured syntactic features, such as for instance the subject of verb type [36]. The authors of [37, 5] exploit a few global syntactic features, but conclude that such features are useless for dialogue act recognition. However, the features that were used in this work are very limited (POS-tags and the MapTask SRule annotation that indicates the main structure of the utterance, i.e., Declarative, Imperative, Inverted or Wh-question), and the authors acknowledge in the discussion that syntactic features may prove useful when used in a different manner and support this remark with two references. Syntax is a very rich source of information and can certainly neither be reduced to POS-tags nor to SRule. Also, the potential impact of syntactic information highly depends on the chosen integration approach and experimental setup. Hence, given the experimental setup described in [37], neither POS-tags nor dialogue act history bring any improvement, when several other works have concluded differently for such features. We thus propose in the next section other types of syntactic features and a different model and show that syntax might indeed prove itself useful for dialogue act recognition. But let us first support our hypothesis by briefly reviewing a few other papers that also support the use of syntax for both dialogue act recognition and closely related domains.

First, as already shown, word n-grams features, with n greater than 1, do implicitly encode local syntactic relations and are used successfully in most dialogue act recognition systems. But more importantly, a recent work [38] concludes that both dialogue context and syntactic features dramatically improve dialogue act recognition, compared to words only, more precisely from an accuracy of 48.1% to 61.9% when including context and 67.4% when including syntactic features as well. They use in their experiments a Bayesian Network model and their syntactic features are the syntactic class of the predicate, the list of arguments and the presence of a negation. Although this work actually focuses on predicate-argument structures, while our main objective is rather to exploit the full syntactic tree without taking into account any semantic-level information for now, this work supports our claim that syntactic information may prove important for dialogue act recognition. In addition, Zhou et al. employ in [39] three levels of features: 1) word level (unigram, bigram and trigram), 2) syntax level (POS-tags and chunks recognized as Base Noun Phrase (BNP)) and 3) restraint information (word position, utterance length, etc.). Syntactic and semantic relations are acquired by information extraction methods. They obtain 88% of accuracy with a SVM classifier on a Chinese corpus and 65% on the SWBD corpus.

3.1 Contributions

As already stated, several different dialogue act recognition approaches have been proposed in the literature. The main goal of our work presented in [40] is to give a brief overview of these approaches. A short description is thus given for each of them, and is usually complemented by a discussion of their theoretical and practical advantages and drawbacks.

We have discovered that a major concern in the DA recognition domain is that, although a few DA annotation schemes seem now to emerge as standards, the DA tag-sets often have to be adapted to the specifics of a given application. This prevents the deployment of standardized DA databases and evaluation procedures.

The main focus of our review is put on the various kinds of information that can be used to recognize DAs, such as prosody, lexical, etc., and on the types of models proposed so far to capture this information. Combining these information sources tends to appear nowadays as a prerequisite to recognize DAs and our future work is based on the conclusions of this review.

We investigated in [41] a new kind of information for dialogue act recognition, that is the word position in the utterance. In contrast to the previous methods (usually word *n-grams*), this information is global at the sentence level. Intuitively, this information is quite important for this task, as for instance, the word “who” is often at the beginning of sentences for questions, and at other positions for declarative sentences. A novel approach that takes into account this information consists in analyzing the sentence into a syntactic tree, but such analyzers are also known to work with errors in spontaneous speech. Hence, our first work in this field is rather based on statistical methods.

Three original dialogue act recognition methods including sentence structure information are proposed:

- The first *multiscale position* method considers the relative positions in a multiscale tree to smooth the models likelihoods;
- Let p be a random variable representing word position in the utterance W. The second *non-linear merging* method models the dependency between W and p by a non-linear function that includes p;
- The third *best position* method decouples the positions from the lexical identities to maximize the available training corpus.

First, we evaluated these approaches on manually transcribed input sentences. Then, the manual transcription has been replaced by an automatic one obtained by the Czech speech recognizer jLASER [42]. This experiment was realized in order to validate the proposed approaches in real applications that are often based on an automatic speech recognition. The resulting decrease in performances is very small, which confirms the validity of the proposed approaches.

We further extended our work and proposed in [43] another approach that derives several simple as well as more complex syntactic features from the deep sentence parse tree. These features are then used as an input to the Conditional Random Fields.

The parse trees are defined within the dependency framework [44] and are automatically computed on the input sentences. We evaluated this approach in two conditions, when the input sentences are manually or automatically transcribed on the Czech Railways corpus [45], respectively. We have shown that the use of such features indeed significantly improves in both cases the dialogue act recognition performances. Finally, we have studied the robustness of the proposed approach and have shown that, as expected, the most complex syntactic features are also the most sensitive to speech recognition errors.

Hence, given the evidence collected in this work, we conclude that syntactic information is important for dialogue act recognition, as it has already been shown to be relevant for many other natural language processing tasks. The main challenge that remains is to increase its robustness to speech recognition errors, but we expect this challenge to be soon overcome, thanks to the great progresses realized in the automatic parsing community in the recent years.

Selected Papers:

- P. Král, C. Cerisara and J. Klečková, **Lexical Structure for Dialogue Act Recognition**, *in Journal of Multimedia (JMM)*, ISSN: 1796-2048, Volume 2, Issue 3, 2007, pp. 1-8.

- P. Král, C. Cerisara, **Dialogue Act Recognition Approaches**, *in Computing and Informatics*, ISSN: 1335-9150, Volume 29, No 2, 2010, pp. 227-250.

- P. Král, C. Cerisara, **Automatic Dialogue act Recognition with Syntactic Features**, *in Language Resources and Evaluation*, ISSN: 1574-020X, E-ISSN: 1574-0218, doi: 10.1007/s10579-014-9263-6, February 2014, pp. 1-23, Springer.

4 Automatic Punctuation Generation

Automatic speech transcriptions are quite difficult to read, because of recognition errors, but also because of the missing structure of the document and in particular capitalization and punctuation. We focus here on the improvement of automatic generation of one of the most difficult punctuation marks, commas, in Czech and French. Adding this information into speech transcripts will make automatic speech transcriptions more understandable and may also help subsequent automatic processing such as parsing or mining.

Punctuation generation is often based on prosodic (pauses, pitch contours, energy) and lexical (surrounding words, n-grams) features, such as in [46], where full stops, commas and question marks are recovered using a finite state approach that combines lexical n-grams and prosodic features. Both prosodic and lexical features are also combined via a maximum entropy model in [8], where commas are recovered on the Switchboard corpus.

In [47], automatic capitalization is realized along with automatic generation of full stops and commas in Portuguese. Both punctuation marks are detected with a maximum entropy model that exploits acoustic and lexical features.

The authors of [48] exploit a hidden-event n-gram model combined with a prosodic model to recover punctuation marks on the Czech broadcast news corpus.

In [49], a maximum entropy model is also exploited to recover 14 punctuation marks from the Penn Chinese TreeBank. The authors also use syntactic features derived from the manual syntactic annotations.

The authors of [50] focus on the study of comma prediction in English with syntactic features. They have compared three sequence models: Hidden-Event Language Model (HELM), factored-HELM and Conditional Random Fields (CRF). They report that the best results have been obtained with CRF, although CRFs may not scale easily to large databases.

4.1 Contributions

We extend in [51] the work of [50] in the following aspects:

- Design of new syntactic features dedicated to comma recovery and derived from dependency structures;
- Proposing a novel comma generation approach (briefly described next) based on the Conditional Random Fields (CRFs) which use the above proposed features;
- Evaluation of these features and approach on two European languages: Czech and French in comma generation task.

In Czech and French, the available corpora are far from being as large as in English, and scaling is not yet an issue. We have thus decided to base our work on CRF models. Furthermore, considering the relatively limited impact of prosodic features for commas recovery as reported in the literature, only lexical and syntactic features are exploited next. A CRF model is then trained to classify every subsequent word into two classes: the class of words that are followed by a comma, and the class of words without comma. The CRF input features are only local and derived from the current, previous and next words. These features are then concatenated, with special words inserted at sentence boundaries, into a feature stream that is used to train the CRF model.

We show that in both languages, the syntactic features improve the performances largely above significance levels. This supports our conclusions on the importance of syntax for this task and extends them from English to Czech and French languages.

Selected Paper:

- C. Cerisara, P. Král, C. Gardent, **Commas recovery with syntactic features in French and in Czech**, *in Interspeech 2011*, Florence, Italy, 27 - 31 August 2011, ISCA, pp. 1413-1416, ISSN: 1990-9772, ISBN: 978-1-61839-270-1.

5 Automatic Document Classification

Document classification relies usually on supervised machine learning methods that exploit a manually annotated training corpus to train a classifier, which in turn identifies the class of new unlabelled documents. Most approaches are based on the Vector Space Models (VSMs), which mostly represent each document as a vector of all occurring words usually weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

One important issue of this task is that the feature space in VSM has a high dimension which negatively affects the performance of the classifiers. Numerous feature selection/reduction approaches have been proposed in order to solve this problem [52, 53].

In the last years, multi-label document classification [54, 55] becomes a popular research field, because it corresponds usually better to the needs of the real applications than the single-label document classification.

Furthermore, a better document representation may lead to decreasing the feature vector dimension, e.g. using lexical and syntactic features as shown in [56]. Chandrasekar et al. further show in [57] that it is beneficial to use POS-tag filtration in order to represent a document more accurately. The authors of [58] and [59] use a set of linguistic features. Unfortunately, they do not show any impact on the document classification task. However, they conclude that more complex linguistic features may improve the classification score.

More recently, an advanced technique based on Labelled Latent Dirichlet Allocation (L-LDA) [60] has been introduced. L-LDA incorporates supervision by constraining the topic model to use only the topics that correspond to document labels. Principal Component Analysis (PCA) [61] incorporating semantic concepts [62] has been also successfully used for the document classification.

Most of the proposed approaches are focused on English. Unfortunately, only little work about the document classification in other non-mainstream languages, particularly in Czech, exists. Hrala et al. [63] use lemmatization and POS-tag filtering for a precise representation of the Czech documents. The authors further show the performance of three multi-label classification approaches [64].

5.1 Contributions

We present in [65, 66] a set of novel sofisticated features to improve the document classification task. We follow the conclusions of [58, 59] and therefore we are not focusing on the linguistic features. We rather study the impact of semantic and related information sources to improve the classification score.

Our first contribution presented in [65] consists in proposing new features based on the Named Entities (NEs). We believe that NEs, representing usually the semantically richest words, bring important information, which can improve the performance of our document classification system.

We propose and evaluate five novel approaches to employ the information contained in the named entities for the document classification task:

1. Add directly the named entities to the feature vector (which is composed of words or lemmas) as new tokens;

2. Concatenate words related to multiple-word entities to one individual token;

3. Combine (1) and (2);

4. Concatenate words and named entities to one individual token;

5. Replace words related to the named entities by their NEs.

Note that, as far as we know, named entities were never used previously as features for classification of the Czech documents. Moreover, we have not found another work which uses NEs similarly for document classification.

We evaluate these methods on the Czech CTK corpus of the newspaper text documents. The experimental results show that these features do not improve significantly the score over the baseline word-based features. The improvement of the classification error rate is about 0.42% when the best approach is used.

We have further analyzed and compared the confusion matrices of the baseline approach with our proposed methods. This analysis has shown that named entities bring some additional information for document classification. Unfortunately, this information is not sufficient to improve the document classification accuracy significantly.

One important issue in the document classification field is the high dimensionality and insufficient precision of the feature vector. Several feature selection methods and sophisticated language specific features have been proposed. The main drawback of these methods is that they need a significant amount of the annotated data. Furthermore, a complete re-annotation is necessary when the target language is modified. We address these issues in [66] by proposing novel fully unsupervised features. The first feature is based on an unsupervised stemmer, while the other ones introduce semantic information using Latent Dirichlet Allocation (LDA) and semantic spaces (HAL and COALS). These features are further integrated into the multi-label document classification task.

The use of semantic space models (i.e. HAL and COALS) is a very important contribution, because they have not been used for document classification yet. The proposed approaches are evaluated on Czech, as a representative of morphologically rich language. It is noteworthy that, to the best of our knowledge, this evaluation has never been done before.

We have experimentally shown that almost all proposed features significantly improve the document classification score.

Selected Papers:

- P. Král, **Named Entities as new Features for Czech Document Classification**, *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014)*, Kathmandu, Nepal, April 6-12, 2014, Springer, pp. 417-427.
- T. Brychcín, P. Král, **Novel Unsupervised Features for Czech Multi-label Document Classification**, *in 13th Mexican International Conference on Artificial Intelligence (MICAI2014)*, Tuxtla Gutirrez, Chiapas, Mexico, November 1622, 2014, Springer, pp. 70-79, ISBN: 978-3-319-13646-2, doi: 10.1007/978-3-319-13647-9_8.

6 Conclusions & Future Work

The main goal of this habilitation thesis is to show the importance of lexical information, syntax and semantics for three fundamental NLP tasks: dialogue act recognition, punctuation generation and automatic document classification.

The first contribution is the demonstration that global word position is beneficial for dialogue act recognition. We have proposed, evaluated and compared three different approaches for this task. Another scientific contribution consists in proposing novel syntactic features for dialogue act recognition. These features are derived from the whole dependency tree and are further integrated into the dialogue recognition system to improve the recognition results. The proposed approaches were experimentally evaluated on the Czech language. Based on our experiments, we concluded that syntactic information is important for dialogue act recognition, as it has already been shown for many other NLP tasks.

The next contribution is the use of syntax for automatic punctuation generation. We proposed novel syntactic features and integrated them into this task with a CRFs classifier. The validity of the proposed approach was demonstrated on Czech and French languages. We have experimentally shown that in both languages, the syntactic features improve significantly the performances of the punctuation generation system. We have thus demonstrated the importance of syntax in two European languages different from English.

We have further shown the importance of semantics for automatic document classification. We proposed several mainly unsupervised features for this purpose and integrated them into a document classification task. We experimentally evaluated the proposed approaches on the Czech language. The realized experiments have shown that almost all novel features significantly improve the document classification accuracy. We have further implemented an experimental document classification system which is currently deployed in the Czech News Agency platform to simplify the manual processing. This system successfully used the novel proposed features.

Another scientific contribution consists in proposing the novel semi-supervised dependency syntactic parser based on a discriminative graphical model. This parser uses only a few manual rules as constraints instead of some annotated sentences. The parser performance is evaluated on three languages, English, French and Czech, on two domains, newspaper texts and broadcast news transcriptions, and with and without gold part-of-speech tags. The performances of our system are encouraging in all conditions, and match those of related state-of-the-art weakly and semi-supervised systems.

6.1 Future Work

6.1.1 Automatic Dialogue Act Recognition

Our dialogue act recognition approaches were evaluated on the Czech Railways corpus which is limited size and with limited DA number. Therefore, the first perspective consists in evaluating our approaches on larger corpora with more dialogue acts. These corpora will include more languages as for instance French and English.

Our current dialogue act models do not use semantics. Therefore, we would like to improve our methods by including this information. We have good experience with semantic spaces and topic modelling in the document classification task. Therefore, the semantic information will be firstly provided by these models. Then, we would like to explore the other semantic sources.

Another important information that has not been taken into account in this work is dialogue act grammar, which models the most probable sequences of dialogue acts. It is straightforward to use such information in our DA recognition system, but we have not yet done so because it somehow masks the influence of the proposed features we focus on in this work. Indeed, this grammar certainly improves the recognition results.

We have used supervised machine learning methods in all above mentioned dialogue act recognition approaches. Unfortunately, these approaches need significant amount of annotated data. Our last perspective in this research area thus consists in using the approaches with minimal or without supervision. These novel approaches will be more general and much more adaptable to the other applications/languages.

6.1.2 Automatic Punctuation Generation & Syntactic Parsing

Our experiments in the automatic punctuation generation field have been realized on manually transcribed speech recordings. We will extend our work to successfully process automatic speech recognition outputs, with the objective of enriching such transcripts with punctuation. However, this requires first to solve the weakness of the current Czech and French parsers, which are not robust enough to recognition errors.

We will further adapt our weakly syntactic parsing framework for punctuation generation. This requires to define novel syntactic rules to discover the different punctuation marks in the speech transcripts and to adapt the discriminative graphical model. One advantage of this novel semi-supervised approach should be its simple adaptation to the other languages/domains.

6.1.3 Automatic Document Classification

First, we plan to validate our methods on other corpora in different languages with different document labels. We will also consider different document/labels distributions. Due to the unsupervised character of the majority of the proposed features, no additional annotations will be required.

Several work [58, 59] concluded that linguistics features do not have any positive impact on the document classification. Therefore, we did not used these features in this work. However, based on our conclusions in the dialogue act recognition and punctuation generation fields and on the recent work of Moschitti [67], we believe that more sophisticated syntactic features can play also a positive role for document classification. Therefore, we would like to use syntactic information to help in some document classification tasks. For example, it is evident that different persons usually use different syntactic structures. Therefore, syntax should help for classification of the speaker turns.

6.1.4 Combining of Natural Language & Image Processing

We have done with Ph.D. student Ladislav Lenc a significant work in the automatic face recognition field (see [68, 69, 70]) and we have thus also a good know-how in the pattern recognition field. We would like to further focus on the tasks where both information (textual and graphical) play an important role.

Our first scientific challenge consists in designing and implementing a system to search appropriate photographs for the newspaper articles from a database of images associated with a short description. One solution is to use only the text data and NLP methods. However, we believe that graphical information contained in the images can improve the matching results. Therefore, we would like to adapt and employ our previously developed pattern matching techniques and combine them with the results of the natural language processing. Note that this research is motivated by the needs of the Czech News Agency to implement a commercial image retrieval application for automatic association of the newspaper articles with suitable photographs.

References

[1] Daniel Jurafsky and H James, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2000, ISBN: 978-0131873216.

[2] Jeffrey S Gruber, *Lexical structures in syntax and semantics*, volume 25, North-Holland publishing company Amsterdam, 1976.

[3] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA, May 1999.

[4] Ralph Grishman, Information extraction: Techniques and challenges, In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27, Springer, 1997.

[5] B. Di Eugenio, Z. Xie, and R. Serafin, Dialogue act classification, higher order dialogue structure, and instance-based learning, *Journal of Discourse and Dialogue Research*, 1(2):1–24, July 2010.

[6] Y.W. Wong and R.J. Mooney, Learning for semantic parsing with statistical machine translation, In *HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference*, pages 439–446, 2006.

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

[8] Jing Huang and Geoffrey Zweig, Maximum entropy model for punctuation annotation from speech, In *7th International Conference on Spoken Language Processing(ICSLP)*, pages 917–920, Denver, Colorado, USA, 16-20 September 2002.

[9] Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Vidová-Hladká, The Prague Dependency Treebank: A Three-Level Annotation Scenario, In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127, Amsterdam: Kluwer, 2000.

[10] Dan Klein, *The unsupervised learning of natural language structure*, PhD thesis, Stanford University, 2005.

[11] W. P. Headden III, M. Johnson, and D. McClosky, Improving unsupervised dependency parsing with richer contexts and smoothing, In *Proc. NAACL*, 2009.

[12] V. I. Spitkovsky, H. Alshawi, A. X. Chang, and Jurafskyn D., Unsupervised dependency parsing without gold part-of-speech tags, In *Proc. EMNLP*, 2011.

[13] J. Graça, K. Ganchev, and B. Taskar, Expectation maximization and posterior constraints, In *Proc. NIPS*, 2007.

[14] J. Gillenwater, K. Ganchev, J. Graça, F. Pereira, and B. Taskar, Posterior sparsity in unsupervised dependency parsing, *Journal of Machine Learning Research*, 12:455–490, February 2011.

[15] Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson, Using universal linguistic knowledge to guide grammar induction, In *Proc. EMNLP*, pages 1234–1244, 2010.

[16] H. Daumé III, Unsupervised search-based structured prediction, In *Proc. ICML*, Montreal, Canada, 2009.

[17] M. S. Rasooli and H. Faili, Fast unsupervised dependency parsing with arc-standard transitions, In *Proc. EACL*, 2012.

[18] C. Cerisara, A. Lorenzo, and P. Král, Weakly supervised parsing with rules, In *Interspeech 2013*, pages 2192–2196, Lyon, France, 25-29 August 2013, ISCA, ISBN: 978-1-62993-443-3.

[19] G. Druck, K. Ganchev, and J. Graça, Rich prior knowledge in learning for natural language processing, In *ACL tutorial*, 2011.

[20] B. Settles and X. Zhu, Behavioral factors in interactive training of text classifiers, In *short paper, Proc. NAACL*, pages 563–567, 2012.

[21] A. Stolcke *et al.*, Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech, In *Computational Linguistics*, volume 26, pages 339–373, 2000.

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

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

[24] 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*, pages 217–229, 1996.

[25] 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.

[26] 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*, pages 88–94, Philadelphia, USA, July 2002.

[27] G. Ji and J. Bilmes, Dialog Act Tagging Using Graphical Models, In *Proc. ICASSP*, volume 1, pages 33–36, Philadelphia, USA, March 2005.

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

[29] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, *Language and Speech*, volume 41 of *Special Double Issue on Prosody and Conversation*, chapter Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?, pages 439–487, 1998.

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

[31] J. Orkin and D. Roy, Semi-automated dialogue act classification for situated social agents in games, In *Proc. of the Agents for Games And Simulations Workshop at the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Toronto, Canada, 2010.

[32] 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*, pages 85–94, Prague, Czech Republic, April 1997.

[33] S. Joty, G. Carenini, and C.-Y. Lin, Unsupervised approaches for dialog act modeling of asynchronous conversations, In *Proc. IJCAI*, Barcelona, Spain, July 2011.

[34] N. Crook, R. Granell, and S. Pulman, Unsupervised classification of dialogue acts using a dirichlet process mixture model, In *Proc. of the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL)*, pages 241–348, 2009.

[35] Daan Verbree, Rutger Rienks, and Dirk Heylen, Dialog-act tagging using smart feature selection; results on multiple corpora, In *The first International IEEE Workshop on Spoken Language Technology (SLT)*, Palm Beach, Aruba, 2006.

[36] T. Andernach, A Machine Learning Approach to the Classification of Dialogue Utterances, In *NeMLaP-2*, Ankara, Turkey, July 1996.

[37] R. Serafin and B. Di Eugenio, LSA: Extending latent semantic analysis with features for dialogue act classification, In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Spain, 2004.

[38] Tina Klüwer, Hans Uszkoreit, and Feiyu Xu, Using syntactic and semantic based relations for dialogue act recognition, In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 570–578, Stroudsburg, PA, USA, 2010, Association for Computational Linguistics, URL `http://portal.acm.org/citation.cfm?id=1944566.1944631`.

[39] Keyan Zhou and Chengqing Zong, Dialog-act recognition using discourse and sentence structure information, In *Proceedings of the 2009 International Conference on Asian Language Processing*, IALP '09, pages 11–16, Washington, DC, USA, 2009, IEEE Computer Society, ISBN: 978-0-7695-3904-1.

[40] P. Král and C. Cerisara, Dialogue act recognition approaches, *Computing and Informatics*, 29(2):227–250, 2010.

[41] P. Král, C. Cerisara, and J. Klečková, Lexical structure for dialogue act recognition, *Journal of Multimedia (JMM)*, 2(3):1–8, 2007, ISSN: 1796-2048.

[42] Tomáš Pavelka and Kamil Ekštein, JLASER: An automatic speech recognizer written in Java, In *XII International Conference Speech and Computer (SPECOM'2007)*, pages 165–169, Moscow, Russia, 2007.

[43] P. Král and C. Cerisara, Automatic dialogue act recognition with syntactic features, *Language Resources and Evaluation*, 48(3):419–441, 8 February 2014, ISSN: 1574-020X, doi: 10.1007/s10579-014-9263-6.

[44] Eva Hajičová, Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus, pages 57–78, 2000.

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

[46] H. Christensen, Y. Gotoh, and S. Renals, Punctuation Annotation Using Statistical Prosody Models, In *ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 35–40, 2001.

[47] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news, *Speech Communication*, 50: 847–862, 2008.

[48] J. Kolář, J. Švec, and J. Psutka, Automatic punctuation annotation in Czech broadcast news speech, In *Proc. SPECOM*, pages 319–325, Saint-Petersburg, 2004, SPIIRAS.

[49] Y. Guo, H. Wang, and J. van Genabith, A linguistically inspired statistical model for Chinese punctuation generation, *ACM Transactions on Asian Language Information Processing*, 9(2):27, 2010.

[50] B. Favre, D. Hakkani-Tür, and E. Shriberg, Syntactically-informed models for comma prediction, In *ICASSP'09*, pages 4697–4700, Taipei, Taiwan, April 2009.

[51] C. Cerisara, P. Král, and C. Gardent, Commas recovery with syntactic features in French and in Czech, In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 1413–1416, Florence, Italy, 28-31 August 2011, ISCA, ISBN: 978-1-61839-270-1.

[52] George Forman, An extensive empirical study of feature selection metrics for text classification, *The Journal of Machine Learning Research*, 3:1289–1305, 2003.

[53] Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi, Experiments on the use of feature selection and negative evidence in automated text categorization, In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '00, pages 59–68, London, UK, UK, 2000, Springer-Verlag, ISBN: 3-540-41023-6, URL `http://dl.acm.org/citation.cfm?id=646633.699638`.

[54] Grigorios Tsoumakas and Ioannis Katakis, Multi-label classification: An overview, *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[55] X. Zhu, *Semi-supervised learning with graphs*, PhD thesis, Carnegie Mellon University, May 2005.

[56] Chul Su Lim, Kong Joo Lee, and Gil Chang Kim, Multiple sets of features for automatic genre classification of web documents, *Information Processing and Management*, 41(5):1263 – 1276, 2005, ISSN: 0306-4573, doi: 10.1016/j.ipm.2004.06.004.

[57] Raman Chandrasekar and Bangalore Srinivas, Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging, 1996.

[58] Alessandro Moschitti and Roberto Basili, Complex linguistic features for text classification: A comprehensive study, In Sharon McDonald and John Tait, editors, *Advances in Information Retrieval*, volume 2997 of *Lecture Notes in Computer Science*, pages 181–196, Springer Berlin Heidelberg, 2004, ISBN: 978-3-540-21382-6, doi: 10.1007/978-3-540-24752-4_14.

[59] Alex KS Wong, John WT Lee, and Daniel S Yeung, Using complex linguistic features in context-sensitive text classification techniques, In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 5, pages 3183–3188, IEEE, 2005.

[60] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning, Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009, Association for Computational Linguistics, ISBN: 978-1-932432-59-6.

[61] J. C. Gomez and M-F. Moens, Pca document reconstruction for email classification, *Computer Statistics and Data Analysis*, 56(3):741–751, 2012.

[62] J. Yun, L. Jing, Yu J., and H. Huang, A multi-layer text classification framework based on two-level representation model, *Expert Systems with Applications*, 39(2):2035–2046, 2012.

[63] Michal Hrala and Pavel Král, Evaluation of the Document Classification Approaches, In *8th International Conference on Computer Recognition Systems (CORES 2013)*, pages 877–885, Milkow, Poland, 27-29 May 2013, Springer.

[64] M. Hrala and P. Král, Multi-label document classification in Czech, In *16th International conference on Text, Speech and Dialogue (TSD 2013)*, pages 343–351, Pilsen, Czech Republic, 1-5 September 2013, Springer, ISBN: 978-3-642-40584, doi: 10.1007/978-3-642-40585-3_44.

[65] P. Král, Named entities as new features for Czech document classification, In *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014)*, volume 8404 LNCS, pages 417–427, Kathmandu, Nepal, 6-12 April 2014, ISBN: 978-3-642-54902-1, doi: 10.1007/978-3-642-54903-8_35.

[66] T. Brychcín and P. Král, Novel unsupervised features for Czech multi-label document classification, In *13th Mexican International Conference on Artificial Intelligence (MICAI 2014)*, pages 70–79, Tuxtla Gutierrez, Chiapas, Mexic, 16-22 November 2014, Springer, ISBN: 978-3-319-13646-2, doi: 10.1007/978-3-319-13647-9_8.

[67] Alessandro Moschitti, Kernel methods, syntax and semantics for relational text categorization, In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 253–262, ACM, 2008.

[68] Ladislav Lenc and Pavel Král, Automatically detected feature positions for LBP based face recognition, In *Artificial Intelligence Applications and Innovations*, pages 246–255, Springer, 2014.

[69] Ladislav Lenc and Pavel Král, Two-step supervised confidence measure for automatic face recognition, In *24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014)*, pages 1–6, Reims, France, 21-24 September 2014, IEEE, ISBN: 978-1-4799-3693-6.

[70] Ladislav Lenc and Pavel Král, Automatic face recognition system based on the SIFT features, *Computers and Electrical Engineering*, 2015, accepted.

Author's Selected Papers

The author's selected papers referenced previously are included below in their published form, in a respective order.

Weakly supervised parsing with rules

C. Cerisara[1], *A. Lorenzo*[1] *and P. Kral*[2,3]

[1]LORIA-UMR7503, Nancy, France
[2]Dept of Computer Science and Engineering, Univ. of West Bohemia, Plzeň, Czech Republic
[3]NTIS - New Technologies for the Information Society, Univ. of West Bohemia, Plzeň, Czech Rep.

cerisara@loria.fr, alelorenzo@gmail.com, pkral@kiv.zcu.cz

Abstract

This work proposes a new research direction to address the lack of structures in traditional n-gram models. It is based on a weakly supervised dependency parser that can model speech syntax without relying on any annotated training corpus. Labeled data is replaced by a few hand-crafted rules that encode basic syntactic knowledge. Bayesian inference then samples the rules, disambiguating and combining them to create complex tree structures that maximize a discriminative model's posterior on a target unlabeled corpus. This posterior encodes sparse selectional preferences between a head word and its dependents. The model is evaluated on English and Czech newspaper texts, and is then validated on French broadcast news transcriptions.

Index Terms: speech parsing, unsupervised training, inference

1. Introduction

N-gram models have several well-known theoretical limitations, most notably regarding the fact that they reduce natural language syntax to a small linear context. Despite these limitations, most, if not all, current state-of-the-art automatic speech recognition systems are based on n-gram models. Several interesting alternative have been proposed that try to introduce structure back into such systems. Probably the most well-known is the Structured Language Model of Chelba and Jelinek [1]. However, these approaches have so far not been able to outperform the n-grams by a large-enough margin to make them become the new mainstream type of language models. The reason for this might come, partially, from both the lack of robustness of current parsers to speech recognition errors and the difficulty to accurately model speech syntax, which does not benefit from as large treebanks as written text in many languages. Thus, we believe that a solution to these issues would be to further investigate the weakly supervised machine learning area and how it could be used to replace corpus annotations by other types of knowledge that shall guide the unsupervised training of parsers designed for speech. Our long-term objective is thus to design accurate parsers for speech transcriptions that would not rely on large annotated treebanks and still would be good enough to be integrated into speech recognition systems thus compensating the lack of structures of current n-gram models.

This work describes the first part of this ambitious goal, that is the proposal of a new weakly supervised parser for speech transcripts that exploits a few hand-crafted rules as a substitute to corpus annotations. We first review the litterature about unsupervised training in Section 2, and then describe our model in Section 3. We then validate experimentally the proposed model in Section 5, first on written text corpora for English and Czech, in order to show that the approach can be used to train new models for different languages at a low cost, and finally on French speech transcripts. The integration of this model into a speech recognition system is another challenging task that is left for future work.

2. Related works

Unsupervised parsing aims at automatically producing syntactic trees on top of a raw, unlabeled text corpus. Many amongst the most successful approaches in the field [2, 3, 4, 5, 6] exploit some stochastic process to decompose parents-children dependencies. The parameters of these models are typically trained so as to maximize the sparsity of selectional preferences in the corpus. Although no manual annotations are given, one can argue that some “linguistic” knowledge is nevertheless introduced in the model's definition, for instance in the set of conditional independencies, priors and initial parameters. It has further been shown that adding some kind of knowledge might greatly improve the performances of unsupervised parsers and help to control their convergence and the resulting structures.

Hence, [7] exploit phylogenetic dependencies between human languages, [8] replace standard corpus annotations with a few syntactic prototypes and [9] make use of semantic cues. The posterior regularization framework [10] is often used to integrate constraints during inference, e.g., with a sparsity-inducing bias over unique dependency types [11] or with a few universal rules that are valid across languages [5].

Most of these works rely on a generative Bayesian model because of fundamental theoretical limitations concerning unsupervised training of discriminative models. Nevertheless, as discussed in Section 3.2, using knowledge to constrain inference makes the training of discriminative models possible even without any supervised annotation [12]. Several different approaches have thus been proposed to train discriminative parsers on unlabeled corpora, for instance by transferring dependency grammars from English to other languages within the posterior regularization framework and with discriminative models [13] or by defining preferred dependency constraints within the generalized expectation framework with tree CRFs [14]. Semi-supervised discriminative parsers can also make a very efficient use of even a few annotated sentences, such as in the SEARN paradigm [15, 16]. Our work has also been inspired by the “Constraint-Driven Learning” paradigm, as proposed in [17, 18] and generalized in [19], as well as with [20, 8].

3. Proposed framework

We propose a weakly supervised approach that relies on two main components: a set of rules that generate dependency structures over input sentences annotated with part-of-speech (POS) tags [1], and a model that evaluates the trees that are produced by the rules on some raw text corpus [2]. The rules shall describe, for every possible dependency type (such as subject, object...), the most standard situations in which this relation may occur.

3.1. Rules design process

The proposed framework was originally designed to be used with hand-crafted rules. However, validating the framework only with manually defined rules may leave some doubts about two potentially problematic aspects: first, the difficulty to reproduce our experimental results, because of the subjectivity in the rules design process. Hence, different users will most likely write different rules for the same task; second, excessive tuning to the task, as it is always possible to improve the results by writing more rules, or fine-tuning them.

To address both issues, we propose next to automatically train and extract the rules from a small labeled corpus. Thereafter, we further validate our approach with hand-crafted rules, to support our original motivation that is to develop a parser without any annotated corpus.

3.1.1. Automatically trained rules

In order to automatically extract our set of rules from a small labeled corpus $\mathcal{C}$, we first define the parametric form of every rule as $R(u, w, h, d, s_R, \mathcal{C})$, where R is the rule that creates one and only one dependency arc with label d from word w of sentence u to the head word h of sentence u. The fifth parameter is the score s_R that represents the level of confidence of this rule: the larger s_R is the more chance has the rule to be correct. This score is computed with a Support Vector Machine (SVM) that is trained on $\mathcal{C}$. The SVM uses the same basic features as the ones used in the MATE parser [22]. An example of a feature is the tuple (d,form(h),postag(h),word_order(w,h)). The full set of features used in our classifier includes all the first-order features [3] listed in Table 4 in [22].

At test time, all possible rules that link every pair of words with every possible dependency labels are first built for each input sentence. Then, the score s_R of every rule is computed with the SVM, and all rules which score is below 0 are removed. The inference algorithm described in Section 3.3 is then applied with these rules, just as it is done with the manual rules, with the exception of two minor differences, which both come from the fact that there are much more automatic rules (up to 4000 rules per sentence) than manual rules (up to 40 rules instances per sentence), leading to a much larger search space. We have thus slightly adapted the proposed inference algorithm to accommodate this increased search space by choosing as initial configuration the trees produced by the MATE parser, which has also been trained on $\mathcal{C}$.

The performances of both the initial MATE model and the proposed model with automatically trained rules are shown in Table 1.

3.1.2. Manually designed rules

The proposed framework supports many types of rules, which shall only take as input a sequence of words, check that some preconditions are met in the input, such as the existence of a noun followed by a verb, and output new annotations on the sentence, such as a subject relation between the noun and the verb.

The user has thus a lot of freedom in the types of rules he may write, and even correlated, ambiguous, incomplete and logically inconsistent rules are allowed, thanks to Bayesian inference that filters-out irrelevant rules. For instance, on the one hand, the user could write a single rule that links any word to any other word with any dependency label. This situation reflects the purely unsupervised case [4]. On the other hand, the user could instead write a large set of so precise rules that the correct parse trees can be derived from them without ambiguity, leading to an ideal rule-based deterministic parser, in which case the proposed Bayesian model is useless. The current work rather aims at some intermediate stage, where the user should write one or a few rules per dependency type, which, when combined, lead to relatively ambiguous parses, and where Bayesian inference should take care of resolving ambiguity to find the correct tree. The potential of the proposed method to support such unconstrained and intuitive rules makes the proposed approach unique in the field.

Given these general guidelines, we have implemented a "rule definition language" that extends traditional regular expressions to manipulate tree-like structures. This allows the user to write a simple text file with regular expressions to match some tree or sequence patterns and produce new dependency arcs. When these regular expressions are not expressive enough for the user, he can directly implement the rule interface in Java code, and thus manipulate the dependency tree structure the way he wants. In the following French and English experiments, the rules are defined using both formats, while only regular expressions are used for Czech.

In order to decide which rules he shall write, the user may use an existing annotation guide or examples of annotated sentences. In the following experiments, as it is difficult to master the three English, French and Czech annotation guides, we have rather given the first 50 labeled sentences per language as examples to the rules designer. Although he actually used only a small fraction of these annotations, we have nevertheless compared the resulting system with other semi-supervised approaches that exploit twice as much annotated sentences.

3.2. Scoring model

The second component of the framework is the model, which scores the full set of trees produced by the rules on the corpus. This score reflects two linguistic criteria: the sparsity of lexical preferences and lists of dependents for a given head word [5]. The search algorithm, described in Section 3.3, then looks for the best sequences of rules, one per sentence of the corpus, that maximize the global model's score. Note that the size of the search space depends on the level of ambiguity of the rules set.

In the following, the scoring model is implemented as a discriminative directed graphical model, shown in Figure 1. Classically, generative models are used in unsupervised systems, because discriminative models cannot learn from unlabeled data [6]. However, our model is not purely unsupervised, because of the rules that constrain the values that the latent variables can take, leading to a *weakly supervised* training algorithm.

In Figure 1, R_u is the main latent variable; it is the se-

[1] In the following English and Czech experiments, the gold POS-tags are considered, while in the French experiments, they are automatically computed with the Treetagger [21]

[2] Initially, the corpus is unlabeled, and the only supervision considered in this work comes from the rules.

[3] i.e., excluding grandparent and siblings features

[4] This case requires to use a generative model instead of the discriminative model proposed in Section 3.2

[5] The DMV model [2] uses similar criteria. Another choice may be to maximize sparsity of recurrent elementary trees, such as in [4]

[6] At least in a theoretically purely unsupervised setting without constraints. See [12] for a discussion and some solutions.

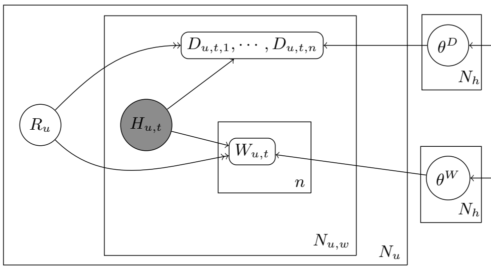


Figure 1: Plate diagram of the scoring model. Shaded nodes are observed and double arrows represent deterministic relations. N_u is the number of utterances in the corpus, $N_{u,w}$ is the number of words in utterance u, N_h is the size of the vocabulary and n is the constant (per head word) maximum number of dependents pre-computed over all possible rules sequences.

quence of rules currently applied onto utterance u. $H_{u,t}$ is the only observed variable; its value is the t^{th} word of utterance u. $D_{u,t} = (D_{u,t,1}, \cdots, D_{u,t,n})$ represents the resulting syntactic frame of H_u, i.e., the ordered list of dependency types governed by H_u and produced by the rules sequence R_u. $W_{u,t}$ encodes lexical preferences, i.e., all words governed by $H_{u,t}$ produced by R_u. The prior of $H_{u,t}$ is assumed uniform, and the only parameters are thus the multinomial parameters θ^D and θ^W, which have respective symmetric Dirichlet priors α^D and α^W. Their concentration parameter is arbitrarily set to 0.001 for both in all experiments.

3.3. Inference

Following standard practice, we perform inference using a Collapsed Gibbs sampler, where the model parameters, θ^D and θ^W, are marginalized out. In each iteration, we want to sample a sequence of rules R_u for each sentence u in turn. In this sampling process, the R_u variable is not decomposed into each of the individual rules that form this sequence, because this may lead to very slow mixing chains, for instance in cases where several rules in the sequence have to be permuted to jump from one mode of the posterior to another. Therefore, besides Gibbs, we still face the challenge of sampling a full sequence of rules per sentence. This may be achieved in several ways. The solution used in the following experiments explores the full search tree of all possible rules permutations in a depth-first manner, which is made possible thanks to the limited length of sentences and to an aggressive pruning based on a topological score that favors projective trees. For longer sentences, R_u sampling may be approximated with an inner loop of Metropolis sampling, which would reject samples that make $P(R_u|R_{-u}, H)$ decrease. Some local hill-climbing search may also be considered to speed-up convergence towards a local optimum, eventually with multiple random restarts or simulated annealing.

4. Experimental validation

All experiments exploit the very same model, with the same α parameters, but of course with different rules sets. The validation procedure consists of first removing the punctuation [7] from all sentences of the corpus, as it is common in the other works we compare to; second, filtering out all sentences that are strictly longer than 10 (for English and Czech) or 15 words (for French); third, initializing the dependency trees by applying all applicable rules in a random order [8]; fourth, running 5000 iterations of Gibbs sampling. The corpus trees that give the highest log-posterior probability are chosen, and the "test" subset of these trees is compared against the gold corpus to compute the standard CoNLL Labeled (LAS) and Unlabeled (UAS) Attachment Scores metrics. The LAS is the ratio of correct head attachments with correct dependency label, while the UAS is the accuracy of head attachments, independently of their labels.

In order to limit the issue of data sparsity during inference, $W_{u,t}$ takes as value the inflected form for words that occur more than 50 times in the corpus and their POS-tag otherwise; the domain of $D_{u,t}$ also only considers a few amongst all possible dependency types: (OBJ, NMOD, PRD, SBJ, VC) for English, (Sb, Obj, Pred, AuxV and AuxT) for Czech, and (AUX, DET, OBJ, SUJ and POBJ) for French.

4.1. English evaluations

Our validation corpus for English is derived from the Penn Treebank [23], after removing all punctuation marks and filtering out all sentences that are strictly longer than 10 words, leading to the standard WSJ10 corpus. The LAS and UAS are reported on Section 23 of this corpus. The first 100 sentences of Sections 2 to 21 are extracted and used with their gold dependency tree to train the MATE parser and to extract our set of automatic rules (see 3.1.1). The rest of Sections 2 to 21 are merged with Section 23 to perform inference, after all dependency trees have been deleted from this corpus.

The last four rows in Table 1 report performances respectively for (i) a baseline that applies the manual rules in a random order; (ii) the proposed system with manual rules; (iii) a baseline formed by the supervised MATE parser trained on the first 100 sentences; (iv) the proposed system with automatic rules trained on the same first 100 sentences and combined with Bayesian inference.

	UAS [%]	LAS [%]
DMV (no rules)	47.1	-
Improved DMV (Headden)	68.8	-
TSG-DMV (Cohn)	66.4	-
Phylogenetic (Berg-Kirkpatrick)	62.3	-
Posterior regularization (Druck)	61.3	-
Post. reg. Universal rules (Naseem)	71.9	-
Post. reg. Collins rules (Naseem)	73.8	-
SEARN 10 sentences (Daumé)	$\sim$ 72	-
SEARN 100 sentences (Daumé)	$\sim$ 75	-
SEARN 1000 sentences (Daumé)	$\sim$ 78	-
Random rules order (10 runs avg)	71.1	50.7
Bayesian inference (5000 iters)	**75.6**	57.0
MATE parser trained on 100 sent.	73.4	64.9
Bayes. inf. with automatic rules	75.0	**65.9**

Table 1: Dep. parsers on the WSJ10 corpus. The confidence interval is $\pm 0.4\%$.

The state-of-the-art results presented in Table 1 come from

[7] Arcs below a punctuation mark are recursively moved up to rather attach to the first head word that is not a punctuation.

[8] Or, for the automatically trained rules, with the trees produced by the MATE parser that has been trained on the first 100 labeled sentences of the training corpus

DMV [2], Improved DMV [3], TSG-DMV [4], Phylogenetic [7], Posterior regularization [14], Post. reg. Universal and Collins rules [5] and SEARN [15].

22 rules have been used in these experiments[9]. This experiment validates the proposed approach on a standard English corpus, and shows that it obtains good results as compared to the state-of-the-art. Our weakly supervised model obtains the best UAS scores with 22 manual rules only, while the best LAS scores are obtained by our model with rules automatically extracted from 100 annotated sentences. The large difference in LAS scores between manual and automatic rules is due to the limited number of manual rules, which cover only 19 out of the 44 dependency types in the WSJ10.

4.2. Czech evaluations

Our Czech evaluation corpus is extracted from the training part of the Prague Dependency Treebank(PDT) [24], with punctuations removed and sentences longer than 10 words filtered out. The remaining corpus, composed of 140 Kwords in 24.5 Ksentences, has further been split into a training (80%) and a test part (20%) in order to match the experimental conditions in [11]. Bayesian inference is realized on the joint train and test corpus, and performances are computed on the test part only. The PDT contains about 2% of non-projective arcs. Although our pruning strategy favors projective trees (see Section 3.3), it does allow crossing dependencies and the rules provide enough constraints to output non-projective trees. Hence, we have observed 2.5% of non-projective arcs in the trees produced by our model. Table 2 reports the obtained results and compare them with the system described in [11], which is an unsupervised DMV-based approach that is trained with additional constraints for dependency type sparsity. The proposed system gives very competitive results with only 14 simple rules.

System	UAS [%]	LAS [%]
DMV (EM algorithm, no rules)	29.6	-
E-DMV (EM-(3,3))	48.9	-
Posterior regularization (Gillenwater)	55.5	-
Random rules order (10 runs avg)	57.3	48.1
Bayesian inference (5000 iters)	58.8	49.4

Table 2: Dep. parsers on the Czech PDT corpus. The confidence interval is $\pm 0.57\%$.

4.3. French evaluations

Although previous semi-supervised parsers have been proposed for French written texts [25], there is no semi-supervised state-of-the-art results to compare with for French broadcast news. We thus compare in Table 3 the proposed model with the supervised MATE parser trained on the 50,000 words that form the training corpus of the Ester Treebank [26]. The Ester Treebank is the only corpus available annotated with dependencies and composed of broadcast news manual speech transcriptions in French. After filtering out all utterances longer than 15 words, the total corpus size on which Bayesian inference is applied is 16,000 words, while the gold contains 1309 words, which leads to a much larger statistical confidence interval than in English.

Although our model's performances are still well below those of a fully trained supervised parser, they give encouraging results without relying on any annotated corpus. We further expect better results by dedicating more time to writing new rules.

System	UAS [%]	LAS [%]
Random rules order (10 runs avg)	61.2	57.0
Bayesian inference (5000 iters)	67.2	62.6
Supervised **MATE**	83.3	78.2

Table 3: Experimental results on the French broadcast news corpus. The confidence interval is $\pm 2.48\%$.

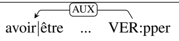
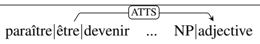
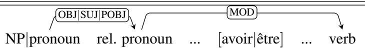
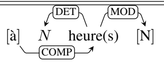
	Rule
Root	Any verb or NP head can be the root of the utterance.
DET	Link any determiner (or number) to the NP head with DET.
AUX	
ATTS	
OBJ	Link with OBJ any NP head or VER:infi or relative pronoun to the preceding verb, or any personal pronoun to the following verb.
SUJ	Link any pronoun or NP head to the next verb with SUJ
COMP	Link any NP head to the preceding preposition, or any verb to the preceding conjunction with COMP
MOD	Link any adverb to the closest adverb, verb, or adjective with MOD
REF	Link any *se* or *s'* to the following verb with REF
DUMMY	Link any *y* to the following verb with DUMMY
rel.	 OBJ|SUJ|POBJ MOD NP|pronoun rel. pronoun ... [avoir|être] ... verb
PP	Link any preposition to the preceding verb with POBJ or MOD, or to the preceding NP head with MOD
time	 and link *à* to the closest verb or noun with MOD
proper names	Link any NAM to the immediately preceding NAM with MOD

Table 4: Rules set for French broadcast news

5. Conclusions

We have proposed a weakly supervised parser that may be used, in a future work, to leverage traditional n-grams with structured dependencies. The integration of this model into a speech recognizer is not described here, but we plan to use it as an additional nbest rescoring pass to start with. This work focuses on the definition and training of the model, which is realized with Bayesian inference on a raw unlabeled corpus. Hand-crafted rules act as constraints to guide inference towards the most plausible solutions from the point of view of the target domain, and especially speech transcripts. To the best of our knowledge, the proposed approach is the first one that includes the rules as latent variables in a discriminative model for parsing, which allows to precisely define their influence on the other meaningful model's variables. Furthermore, the rules are sampled just like any other latent variable, hence giving the model the possibility to ignore some badly defined constraints and increasing its robustness to user mistakes. Another advantage is the high degree of freedom that the user has to write the rules, and the fact that our framework supports both generative and discriminative models. The proposed model is evaluated on three languages, English, French and Czech, on two domains, newspaper texts and broadcast news transcriptions, and with and without gold part-of-speech tags. The model's performances are encouraging across all conditions, and match those of related state-of-the-art weakly and semi-supervised systems.

6. Acknowledgment

This work has been partly supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

[9] We only list next the French rules, because of paper size. English and Czech rules will be made available with the software

7. References

[1] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, pp. 283–332, 2000.

[2] D. Klein, "The unsupervised learning of natural language structure," Ph.D. dissertation, Stanford University, 2005.

[3] W. P. Headden III, M. Johnson, and D. McClosky, "Improving unsupervised dependency parsing with richer contexts and smoothing," in *Proc. NAACL*, 2009.

[4] P. Blunsom and T. Cohn, "Unsupervised induction of tree substitution grammars for dependency parsing," in *Proc. EMNLP*, 2010.

[5] T. Naseem, H. Chen, R. Barzilay, and M. Johnson, "Using universal linguistic knowledge to guide grammar induction," in *Proc. EMNLP*. ACL, 2010, pp. 1234–1244.

[6] V. I. Spitkovsky, H. Alshawi, A. X. Chang, and J. D., "Unsupervised dependency parsing without gold part-of-speech tags," in *Proc. EMNLP*, 2011.

[7] T. Berg-Kirkpatrick and D. Klein, "Phylogenetic grammar induction," in *Proc. ACL*, Uppsala, Sweden, Jul. 2010, pp. 1288–1297.

[8] P. Boonkwan and M. Steedman, "Grammar induction from text using small syntactic prototypes," in *Proc. IJCNLP*, Chiang Mai, Thailand, Nov. 2011, pp. 438–446.

[9] T. Naseem and R. Barzilay, "Using semantic cues to learn syntax," in *AAAI*, W. Burgard and D. Roth, Eds. AAAI Press, 2011.

[10] J. Graça, K. Ganchev, and B. Taskar, "Expectation maximization and posterior constraints," in *Proc. NIPS*, 2007.

[11] J. Gillenwater, K. Ganchev, J. Graça, F. Pereira, and B. Taskar, "Posterior sparsity in unsupervised dependency parsing," *Journal of Machine Learning Research*, vol. 12, pp. 455–490, Feb. 2011.

[12] H. Daumé III, "Semi-supervised or semi-unsupervised?" in *Proc. NAACL Wokshop on Semi-supervised Learning for NLP*, 2009.

[13] K. Ganchev, J. Gillenwater, and B. Taskar, "Dependency grammar induction via bitext projection constraints," in *Proc. ACL*, 2009.

[14] G. Druck, G. Mann, and A. McCallum, "Semi-supervised learning of dependency parsers using Generalized Expectation criteria," in *Proc. ACL*, Suntec, Singapore, Aug. 2009, pp. 360–368.

[15] H. Daumé III, "Unsupervised search-based structured prediction," in *Proc. ICML*, Montreal, Canada, 2009.

[16] M. S. Rasooli and H. Faili, "Fast unsupervised dependency parsing with arc-standard transitions," in *Proc. EACL*, 2012.

[17] M. Chang, L. Ratinov, and D. Roth, "Guiding semi-supervision with constraint-driven learning," in *ACL*. Prague, Czech Republic: Association for Computational Linguistics, 6 2007, pp. 280–287. [Online]. Available: http://cogcomp.cs.illinois.edu/papers/ChangRaRo07.pdf

[18] S. Singh, L. Yao, S. Riedel, and A. McCallum, "Constraint-driven rank-based learning for information extraction," in *Proc. NAACL*, 2010, pp. 729–732.

[19] P. Liang, M. I. Jordan, and D. Klein, "Learning from measurements in exponential families," in *Proc. ICML*, 2009.

[20] A. Haghighi and D. Klein, "Prototype-driven grammar induction," in *Proc. ACL*, Sydney, Jul. 2006, pp. 881–888.

[21] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proc. International Conference on New Methods in Language Processing*, 1994, pp. 44–49.

[22] B. Bohnet, "Top accuracy and fast dependency parsing is not a contradiction," in *Proc. International Conference on Computational Linguistics*, Beijing, China, 2010.

[23] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[24] J. Hajič, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*. Charles University Press, 1999, ch. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank, pp. 106–132.

[25] M. Candito, B. Crabbé, and D. Seddah, "On statistical parsing of french with supervised and semi-supervised strategies," in *Proc. EACL*, 2009.

[26] C. Cerisara, C. Gardent, and C. Anderson, "Building and exploiting a dependency treebank for french radio broadcasts," in *Proc. Intl Workshop on Treebanks and Linguistic Theories (TLT)*, Tartu, Estonia, Dec. 2010.

Computing and Informatics, Vol. 29, 2010, 227–250

DIALOGUE ACT RECOGNITION APPROACHES

Pavel Král

Department of Computer Science and Engineering
University of West Bohemia, Plzeň, Czech Republic
e-mail: pkral@kiv.zcu.cz

Christophe Cerisara

LORIA UMR 7503
BP 239 – 54506 Vandoeuvre, France
e-mail: cerisara@loria.fr

Manuscript received 22 July 2008; revised 22 July 2009
Communicated by Jozef Kelemen

Abstract. This paper deals with automatic dialogue act (DA) recognition. Dialogue acts are sentence-level units that represent states of a dialogue, such as questions, statements, hesitations, etc. The knowledge of dialogue act realizations in a discourse or dialogue is part of the speech understanding and dialogue analysis process. It is of great importance for many applications: dialogue systems, speech recognition, automatic machine translation, etc. The main goal of this paper is to study the existing works about DA recognition and to discuss their respective advantages and drawbacks. A major concern in the DA recognition domain is that, although a few DA annotation schemes seem now to emerge as standards, most of the time, these DA tag-sets have to be adapted to the specificities of a given application, which prevents the deployment of standardized DA databases and evaluation procedures. The focus of this review is put on the various kinds of information that can be used to recognize DAs, such as prosody, lexical, etc., and on the types of models proposed so far to capture this information. Combining these information sources tends to appear nowadays as a prerequisite to recognize DAs.

Keywords: Bayesian approaches, dialogue act, lexical information, prosody, syntactic information

1 INTRODUCTION

Modeling and automatically identifying the structure of spontaneous dialogues is very important to better interpret and understand them. The precise modeling of spontaneous dialogues is still an open issue, but several specific characteristics of dialogues 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.

Dialogue acts can also be used in the context of Spoken Language Understanding. In such systems, dialogue acts are defined much more precisely, but are also application-dependent. Hence, Jeong et al. define in [2] a dialogue act as a domain-dependent intent, such as "Show Flight" or "Search Program" respectively in the flight reservation and electronic program guide domains.

Table 1 shows an example of the beginning of a dialogue between two friends, with Peter (A) calling Michal (B) on the phone. The corresponding DA labels are also shown. Each utterance is labeled with a unique DA.

Speaker	Dialogue Act	English
A	Conventional-opening	Hallo!?
B	Conventional-opening	Hi Peter!
B	Statement	It's me, Michael.
B	Question	How are you?
A	Conventional-opening	Hello Michael!
A	Statement	Very well.
A	Question	And you?
B	Statement	I'm well too.

Table 1. Example of the beginning of a dialogue between persons A and B in English with the corresponding DA labels

1.1 Applications

There are many applications of automatic dialogue acts detection. We mention here only the most important ones: dialogue systems, machine translation, Automatic Speech Recognition (ASR), topic identification [3] and animation of talking head.

In dialogue systems, DAs can be used to recognize the intention of the user, for instance when the user is requesting some information and is waiting for it, or when the system is trying to interpret the feedback from the user. An example of a dialogue management system that uses DA classification is the *VERBMOBIL* [4] system.

In machine translation, dialogue acts can be useful to choose the best solution when several translations are available. In particular, the grammatical form of an utterance may depend on its intention.

Automatic detection of dialogue acts can be used in ASR to increase the word recognition accuracy, as shown for example in [5]. In this work, a different language model is applied during recognition depending on the actual DA.

A talking head is a model of the human head that reproduces the speech of a speaker in real time. It may also render facial expressions that are relevant to the current state of the discourse. Exploiting DA recognition in this context might make the animation more natural, for example by raising the eyebrows when a question is asked. Another easier option is to show this complementary information with symbols and colors near the head.

1.2 Objectives

Recognizing dialogue acts thus can be seen as the first level of dialogue understanding and is an important clue for applications, as it has been shown in the previous section. Several different dialogue act recognition approaches have been proposed in the literature. The main goal of this paper is to give a brief overview of these approaches. A short description is thus given for each of them, and is most often complemented by a discussion of their theoretical and practical advantages and drawbacks.

1.3 Paper Structure

This paper is organized as follows. The first section presents an introduction about the importance of dialogue act recognition with its main applications and objectives. Section 2 briefly describes the task of dialogue act recognition. Sections 3 and 4 describe the most common existing DA recognition approaches. The last section summarizes and discusses them altogether.

2 DIALOGUE ACT RECOGNITION

The first step to implement a dialogue act recognition system consists in defining the set of DAs labels that is relevant for the task. Then, informative features have to be computed from the speech signal and DA models are trained on these features. The segmentation of the dialogue into utterances may be carried out independently from DA recognition, or alternatively realized during the recognition step with joint DA recognition and segmentation models.

2.1 Dialogue Act Tag-set

The DA tag-set definition is an important but difficult step, because it results from a compromise between three conflicting requirements:

1. the DA labels should be generic enough to be useful for different tasks, or at least robust to the unpredictable variability and evolution of the target application;
2. the DA labels must be specific enough to encode detailed and exploitable characteristics of the target task;
3. the DA labels must be clear and easily separable, in order to maximize the agreement between human labelers.

Many different DA tag-sets can be found in the literature, the oldest being reviewed in [6]. Recently, a few of them seem to emerge as a common baseline, from which application-specific DA tags are derived. These are the Dialogue Act Markup in Several Layers (DAMSL) [7], the Switchboard SWBD-DAMSL [8], the Meeting Recorder [9], the VERBMOBIL [10] and the Map-Task [6] DAs tag-sets.

DAMSL was initially designed to be universal. Its annotation scheme is composed of four levels (or dimensions): communicative status, information level, forward looking functions and backward looking functions. Generally, these dimensions are considered as orthogonal and it shall be possible to build examples for any possible combination of them. The communicative status states whether the utterance is uninterpretable, abandoned or it is a self-talk. This feature is not used for most of the utterances. The information level provides an abstract characterization of the content of the utterance. It is composed of four categories: task, task-management, communication-management and other-level. The forward looking functions are organized into a taxonomy, in a similar way as actions in traditional speech act theory. The backward looking functions show the relationship between the current utterance and the previous dialogue acts, such as accepting a proposal or answering the question. DAMSL is composed of 42 DA classes.

SWBD-DAMSL is the adaptation of DAMSL to the domain of telephone conversations. Most of the SWBD-DAMSL labels actually correspond to DAMSL labels. The Switchboard corpus utterances have first been labeled with 220 tags. 130 of those labels that occurred less than 10 times have been clustered, leading to 42 classes.

The Meeting Recorder DA (MRDA) tag-set is based on the SWBD-DAMSL taxonomy. The MRDA corpus contains about 72 hours of naturally occurring multiparty meetings manually-labeled with DAs and adjacency pairs. Meetings involve regions of high speaker overlap, affective variation, complicated interaction structures, abandoned or interrupted utterances, and other interesting turn-taking and discourse-level phenomena. The tags are not organized anymore on a dimensional level (such as DAMSL), but the correspondences are rather listed at the tag level. Each DA is described by one *general* tag, which may be for several DAs completed by one (or more) *specific* tag. A specific tag is used when the utterance cannot be sufficiently characterised by a general tag only. For example, the utterance "Just write it down!" is characterised by the general tag *statement* and by the additional specific tag *command.* MRDA contains 11 general tags and 39 specific tags.

The DA hierarchy in VERBMOBIL is organized as a decision tree. This structure is chosen to facilitate the annotation process and to clarify relationships between different DAs. During the labeling process, the tree is parsed from the root to the leaves, and a decision about the next branch to parse is taken at each node (cf. Figure 1).

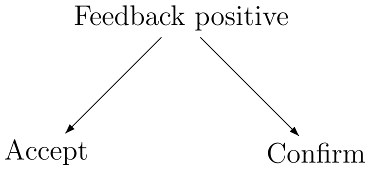


Fig. 1. Part of the VERBMOBIL DAs decision tree hierarchy

42 DAs are defined in VERBMOBIL for German, English and Japanese, with 18 DAs at the illocutionary level.

The DA tags in the Map Task corpus [6] are structured into three levels, the highest modeling *transactions*, where each transaction accomplishes one major step in the speakers' plan. Transactions are then composed of *conversational games*, which model the regularity between questions/answers, statements/denial or acceptance, and so on. Games are finally made up of *conversational moves*, which classify different kinds of games according to their purposes. 19 moves are thus structured hierarchically into a decision-tree that is used to label each DA. For instance, the root of the trees splits into three moves: initiation, response and preparation. Initiation itself is then splitted into command, statement and question, and so on. Moves sequences are then delimited into conversational games, which start with an initiation and ends when that initiation's purpose is either fulfilled or abandoned. Each game is labeled with its purpose, whether it is a top level game or an embedded game, and is delimited in time. Transactions include task description, and are thus application-dependent.

2.2 Dialogue Act Recognition Information

The most important types of information commonly used to recognize dialogue acts are described below.

The first one is *lexical information.* Every utterance is composed of a sequence of words. Generally, the DA of an utterance can be partly deduced from the lists of words that form this utterance. For example, Wh-questions often contains an interrogative word, which rarely occurs in other DA classes. Lexical information is typically captured by words unigrams.

The second one is *syntactic information.* It 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 in dialogue act recognition to model some local syntactic information. Král et al. propose in [11] to further model words position in the utterance in order to also take into account global syntactic information. Another type of syntactic information recently used for DA recognition are "cue phrases", which actually corresponds to 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 dialogue act and on their occurrence frequency [12]. These cue phrases actually correspond to common and typical sequences of words. As they do not model the whole lexical space, one might interpret them in a context of DA detection instead of DA recognition.

Another information is *semantic information.* The DA also depends on the meanings of the utterance and the words that compose it. However, many different definitions of "semantic information" exist, ranging from broad topic categories such as "weather", "sports", down to precise frame-based interpretations, e.g. "show flights from London to Paris on March 12^{th}". The latter is typically used in spoken language understanding applications, where a dialogue act is dependent on a specific pre-defined action [2]. Another kind of semantic information that is used in DA recognition is specific entities, such as named or task entities. For instance, date, place or proper nouns, when they are uttered, may constitute important cues to find out what is the utterance dialogue act [13]. Also, Bangalore et al. use in [14] speaker and task entities as features. They obtain a DA error rate of 38.8 % with 67 dialogue acts adapted from DAMSL on a product ordering task.

Yet another useful information to recognize DAs is *prosody*, and more particularly the melody of the utterance. Usually, questions have an increasing melody at the end of utterance, while statements are often characterised by a slightly decreasing melody.

The last information mentioned here is the *context* of each DA. Hence, any DA depends on the previous (and next) DAs, the most important context being the previous one. For example, a "Yes" or "No" answer is most likely to follow a *Yes/no question.* The sequence of DAs is also called the *dialogue history.*

We focus next on the three following information sources, which are the most commonly used in application-independent DA recognition systems [15, 16]:

- Lexical (and syntactic) information
- Prosodic information
- Dialogue history

2.2.1 Lexical Information

Lexical and syntactic features can be derived from the word sequence in the dialogue. The first broad group of DA recognition approaches that uses this type of features is based on the assumption that different dialogue acts are generally composed of sequences of different words.

The correspondence between DAs and words sequences is usually represented either by n-grams, Naive Bayes, Hidden Markov Models, Bayesian Networks, etc. (see Section 3), or Non-Bayesian approaches, such as Neural Networks, Semantic Classification and Regression Trees, etc. (see Section 4).

2.2.2 Prosodic Information

Most researchers agree on the fact that the lexical/syntactic information is not generally sufficient to explain DAs. Prosodic cues [17] are also related to DA instances.

For example, questions are usually characterized by an increasing melody at the end of the utterance [18], and accepts have usually much more energy than backchannels and acknowledgments [9].

Prosody is successfully used in [19] for French and Vietnamese question detection. Authors exploit the fact that French questions are usually characterized by their intonation curves. The set of prosodic features is derived from the curve of the fundamental frequency (F0). Some features are F0 statistics (Min, Max, Mean, etc.), while other features describe whether F0 is raising or falling. According to the authors, Vietnamese questions and affirmative sentence differ in the F0 contour at the final segment of the sentence, both in register and intensity. They obtain 74 % and 73 % of accuraccy of the French DELOC (telephone meetings) and NESPOLE [20] corpora, respectively. Their question detection accuracy on the Vietnamese VietP corpus is 77 %.

Prosodic features are usually modeled with the same Bayesian or Non-Bayesian methods as used for lexical information.

2.2.3 Dialogue History

The third general type of information used in classical DA recognition systems is the dialogue history. It is defined by the sequence of previous DAs that have been recognized. It may be used to predict the next DA. Different formalisms are employed to model this information: statistical models such as n-grams, Hidden Markov Models (HMMs), Bayesian Networks, etc.

2.3 Segmentation

To recognize DAs, the dialogue must first be segmented into sentence-level units, or utterances [21], where each utterance represents a single DA. Segmentation of the dialogue into such utterances may be carried out separately or realized during the recognition step.

The hidden-event language model has been proposed in [22] to automatically detect utterance boundaries. Its basic principle consists in modeling the joint probability of words and sentence boundaries with an n-gram. The training of the model is realized as in the classical n-gram case with a new token that represents the DA boundary. Shriberg et al. show in [23] that prosodic features give better results than lexical features to segment utterances.

Kolář et al. show in [24] an extension of this approach. They adapt the hidden-event language models to the speaker to improve dialogue act segmentation accuracy. Speaker adaptation is realized by linear combination of the speaker independent and speaker dependent language models. They use ICSI meeting corpus [25].

Ang et al. use in [26] a decision tree that estimates the probability of occurrence of a DA boundary after each word based on the length of the pause between contiguous words of the same speaker, and a bagging classifier that models prosodic attributes. This approach is further combined via an HMM with an hidden-event language model.

The main focus of this review being dialogue act recognition, in the following we will not detail the works about utterance segmentation. Please refer for example to [27] for an overview of this domain.

3 BAYESIAN APPROACHES

The main types of automatic DA recognition approaches proposed in the literature can be broadly classified into Bayesian and Non-Bayesian approaches. Bayesian approaches are presented in this section and Non-Bayesian approaches are described in Section 4.

3.1 Lexical (and Syntactic) N-Gram DA Models

The Bayesian formalism has been the preferred approach in the DA recognition domain for a long time now. For instance, [28] finds the best sequence of dialogue acts $\hat{C}$ by maximizing the *a posteriori* probability $P(C|O)$ over all possible sequences of dialogue acts C as follows:

$$\begin{aligned}\hat{C} &= \arg\max_C P(C|O)\\ &= \arg\max_C \frac{P(C).P(O|C)}{P(O)}\\ &= \arg\max_C P(C).P(O|C).\end{aligned} \tag{1}$$

The most common methods model $P(O|C) = P(W|C)$, where W is the word sequence in the pronounced utterance with statistic models such as n-grams. These methods are based on the observation that different DA classes are composed of distinctive word strings. For example, 92.4 % of the "uh-huh" occur in Backchannels and 88.4 % of the trigrams "<start> do you" occur in yes-no questions [15]. The words order and positions in the utterance may also be considered. A theory of word frequencies, which is the basis for DA modeling from word features, is described in [3].

3.1.1 DA Recognition from Exact Words Transcriptions

The following approach is based on the hypothesis that the words in the utterances are known. Then, Equation 1 becomes:

$$\arg\max_C P(C|W) = \arg\max_C P(C).P(W|C). \tag{2}$$

The "Naive Bayes assumption", which assumes independence between successive words, can be applied and leads to:

$$\arg\max_C P(C).P(W|C) = \arg\max_C P(C).\prod_{i=1}^{T} P(w_i|C). \tag{3}$$

This equation represents the unigram model, also sometimes called the Naive Bayes classifier. In this case, only lexical information is used. Higher order models, such as 2-grams, 3-grams, etc., also take into account some local syntactic information about the dependencies between adjacent words. Because of limited corpus sizes, the use of 4-grams and more complex models is rare.

Reithinger et al. use in [29] unigram and bigram language models for DA recognition on the VERBMOBIL corpus. Their DA recognition rate is about 66 % for German and 74 % for English with 18 dialogue acts. In [30], a naive Bayes n-gram classifier is applied to the English and German languages. The authors obtain a DA recognition rate of 51 % for English and 46 % for German on the NESPOLE corpus. Grau et al. use in [31] the naive Bayes and uniform naive Bayes classifiers with 3-grams. Different smoothing methods (Laplace and Witten Bell) are evaluated. The obtained recognition rate is 66 % on the SWBD-DAMSL corpus with 42 DAs. Ivanovic also uses in [32] the naive Bayes n-grams classifier and obtains about 80 % of recognition rate in the instant messaging chat sessions domain with 12 DAs classes derived from the 42 DAs of DAMSL.

One can further assume that all DA classes are equi-probable, and thus leave the $P(C)$ term out:

$$\hat{C} = \arg\max_C P(W|C). \tag{4}$$

This approach is referred to as the *uniform* naive Bayes classifier in [31].

3.1.2 DA Recognition from Automatic Word Transcription

In many real applications, the exact words transcription is not known. It can be computed approximately from the outputs of an automatic speech recognizer. Let A be a random variable that represents the acoustic information of the speech stream (e.g. spectral features).

The word sequence W is now a hidden variable, and the observation likelihood $P(A|C)$ can be computed as:

$$P(A|C) \quad = \quad \sum_W P(A|W,C).P(W|C) \tag{5}$$

$$= \sum_{W} P(A|W).P(W|C) \tag{6}$$

where C is the DA class and $P(A|W)$ is the observation likelihood computed by the speech recognizer for a given hypothesized word sequence W. Most of the works on Bayesian dialogue act recognition from speech, such as in [15], use this approach and approximate the summation over the k-best words sequence only.

3.2 Dialogue Sequence N-Gram Models

The dialogue history also contains very important information to predict the current DA based on the previous ones. The dialogue history is usually modeled by a statistical discourse grammar, which represents the prior probability $P(C)$ of a DA sequence C.

Let C_τ be a random variable that represents the current dialogue act class at time τ. The dialogue history H is defined as the previous sequence of DAs: $H = (C_1, \dots, C_{\tau-1})$. It is usually reduced to the most recent n DAs: $H = (C_{\tau-n+1}, \dots, C_{\tau-1})$. The most common values for n are 2 and 3, leading to 2-gram and 3-gram models. In order to train such models, the conditional probabilities $P(C_\tau|C_{\tau-n+1}, \dots, C_{\tau-1})$ are computed on a labeled training corpus. *Smoothing* techniques, such as standard back-off methods [33], may also be used to train high-order n-grams. n-grams are successfully used to model dialogue history in [15, 34].

Polygrams are mixtures of n-grams of varying order: n can be chosen arbitrarily large and the probabilities of higher order n-grams are interpolated by lower order ones. They usually give better recognition accuracy than standard n-grams and are shown in [35].

3.3 Hidden Markov Models

Hidden Markov Models can also be used as in [15] to model sequences of dialogue acts. Let O be a random variable that represents the observations and C the sequence of DAs classes. n^{th}-order HMM can be considered, which means that each dialogue act depends on the n previous DAs (in a similar way as for n-grams). Then, each HMM state models one DA and the observations correspond to utterance level features. The transition probabilities are trained on a DA-labeled training corpus.

DA recognition is carried out using some dynamic programming algorithm such as the Viterbi algorithm.

HMMs with word-based and prosodic features are successfully used to model dialogue history in [36]. [5] uses intonation events and tilt features such as: F0 (fall/rise, etc.), energy, duration, etc. She achieves 64 % of accuracy on the DCIEM map task corpus [37] with 12 DA classes. Ries combines in [38] HMMs with neural networks (c.f. Section 4.1). He obtains about 76 % of accuracy on the CallHome Spanish corpus. In [39] language models and modified HMMs are applied on the Switchboard corpus [40] with the SWBD-DAMSL tag-set.

3.4 Bayesian Networks

A Bayesian network is represented by a directed acyclic graph. Nodes and arcs respectively represent random variables and relations (dependencies) between nodes. The topology of the graph models conditional independencies between the random variables. In the following, we do not differentiate dynamic Bayesian networks (with stochastic variables) from static Bayesian networks, as most of our variables are stochastic, and when static Bayesian networks are drawn, they represent an excerpt of a dynamic Bayesian network at a given time slice. The stochastic variables are conditionally dependent of theirs descendants and independent of theirs ascendants.

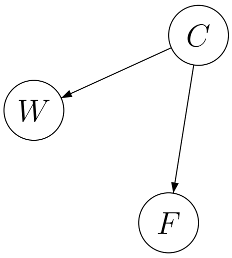


Fig. 2. Example of Bayesian network for dialogue act recognition

An example of Bayesian network for dialogue act recognition is shown in Figure 2. Node C represents the current dialogue act. Utterance features are represented by nodes W (sequence of words in the utterance) and F (prosodic features). The dialogue context is not considered there. The conditional independence assertions of this network allows the following factorization:

$$P(C,W,F) = P(W|C).P(F|C).P(C). \tag{7}$$

In order to build such a network, the network structure (conditional dependencies) and the conditional probability distributions must be defined. The conditional probabilities are trained statistically on a training corpus. The topology of network can be created manually or automatically.

Bayesian networks are successfully used in [41] for dialogue act recognition. In the first experiment reported, three features are used: sentence type (declarative, yes/no question, etc.), subject type (1st/2nd/3rd person) and punctuation (question mark, exclamation mark, comma, etc). The Bayesian network is defined manually. They achieve 44 % of accuracy on the SCHISMA corpus [42]. In the second experiment, a small corpus is derived from the dialogue system used to interact with the navigation agent. Utterances are described by surface level features, mainly keyword-based features. These features are computed automatically for each utterance. Bayesian networks are further generated automatically iteratively, starting from a small hand-labeled DA corpus. This network is used to parse another large corpus, and a new network is generated from this corpus. This approach gives 77 % of accuracy for classification of forward-looking functions (7 classes) and 88 % of accuracy for backward-looking functions (3 classes).

Another application of Bayesian network in dialogue act recognition is shown in [43]. Two types of features are used: utterance features (words in the utterance: w_i) and context features (previous dialogue act: $C_{\tau-1}$). The authors compare two different Bayesian networks to recognize DAs (see Figure 3).

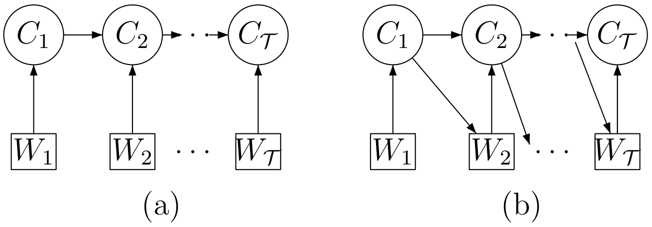


Fig. 3. Two Bayesian networks for dialogue act recognition: C_i represents a single DA, while W_i is a sequence of words

These networks are built manually. In the left model of Figure 3, each dialogue act is recognized from the words of the current utterance and from the previous DA. In the right model of Figure 3, the authors further consider an additional dependency between each word of the utterance and its previous dialogue act (diagonal arcs). They achieve about 64 % precision on a subset of the MRDA corpus and with the reduced DA set size.

Another Bayesian model, the triangular-chain conditional random field, which jointly models dialogue acts and named entities, has been proposed in [2]. This model is shown in Figure 4.

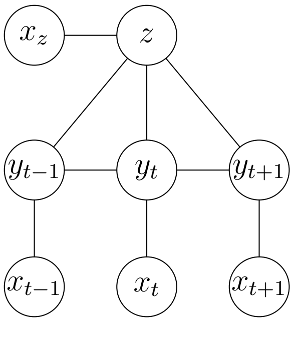


Fig. 4. Triangular-chain Conditional Random Field, from [2]. It is used to jointly model dialogue acts (represented by variables z) and named entities (represented by variables y). Variable x encodes the words sequence.

This joint model is shown to outperform sequential and cascade models, in which dialogue acts are assumed independent of named entities. In the independent approach, DAs are often modeled by a multivariate logistic regression (or maximum entropy classifier)

$$p(z|x) = \frac{1}{Z_z(x)} \exp\left(\sum_k \nu_k h_k(z,x)\right)$$

that maximizes the entropy $h_k(z,x)$, where z is the DA and x the words sequence. Alternatively, the joint model combines both maximum entropy and conditional random fields approaches.

Dynamic Bayesian Network (DBN) have also successfully been used for DA recognition in [44], where a switching DBN combines several partial models and coordinates the DA recognition task. The relation between the sequences of transcribed words and their DA labels is modeled by an interpolated Factored Language Model (FLM), while the dialogue history is represented by a trigram language model. Prosodic features (pitch, energy, etc.) are also used for segmentation. The proposed approach is based on a switching DBN model that alternates between two submodels: an *intra-DA model* that represents a single DA class associated to a words sequence, and an *inter-DA model* that is activated at DA boundaries. A dedicated random variable of these models is used to detect these DA boundaries. The authors obtain about 60 % of DA tagging rate with 15 DA classes on the AMI Meeting Corpus [45].

4 NON-BAYESIAN APPROACHES

Non-Bayesian approaches are also successfully used in the DA recognition domain, but they are not so popular as Bayesian approaches. Examples of such approaches are Neural Networks (NNs), such as Multi-Layer Perceptron (MLP) or Kohonen Networks, Decision Trees, Memory-Based Learning and Transformation-Based Learning.

4.1 Neural Networks

A neural network (NN) [46] is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. It can be used to model complex relationships between inputs and outputs or to find patterns in data.

4.1.1 Multi-Layer Perceptron

One of the most frequently used neural network technique in the DA recognition domain is the multi-layer perceptron (MLP, see Figure 5), which consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and one output layer. The input signal propagates through the network layer-by-layer. An MLP can represent a non linear function.

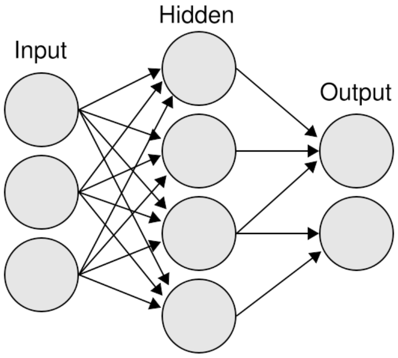


Fig. 5. Example of multi-layer perceptron

Wright describes in [5] an approach with a one-hidden-layer MLP. 54 suprasegmental and duration prosodic features are used as inputs. She achieves 62 % of accuracy on the DCIEM map task corpus [37] with 12 DA classes. Ries successfully uses in [38] an MLP both stand-alone, and in combination with HMMs. He obtains a similar accuracy (about 76 %) on the CallHome Spanish corpus with both setups. Sanchis et al. also use in [47] an MLP to recognize DAs. The features considered are the words of the lexicon restricted to the semantic task (138 inputs=size of the lexicon). The experiments are performed on the Spanish dialogue corpus in the train transport domain (16 DA classes). They achieve about 93 % of accuracy on the text data and about 72 % of accuracy on the recognized speech. Note that this approach may be difficult to apply on a large lexicon. Levin et al. use in [30] a set of binary features to train an MLP. These features are computed automatically by combining grammar-based phrasal parsing and machine learning techniques. They obtain a DA recognition accuracy of about 71 % for English and about 69 % for German on the NESPOLE corpus.

4.1.2 Kohonen Networks

Another type of neural network used in the dialogue act classification domain is Kohonen Networks. A Kohonen network [48], also known as Self-Organizing Map (SOM), defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. A model is associated with each grid node (see Figure 6).

The topology of the SOMs is a single layer feedforward network where the discrete outputs are arranged into a low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. A weight vector with the same dimensionality as the input vectors is attached to every neuron. The number of input dimensions is usually much larger than the output grid dimension. SOMs are mainly used for dimensionality reduction.

The models of the Kohonen network are estimated by the SOM algorithm [49]. A data item is mapped onto the node which model is the most similar to the data item, i.e. has the smallest distance to the data item, based on some metric.

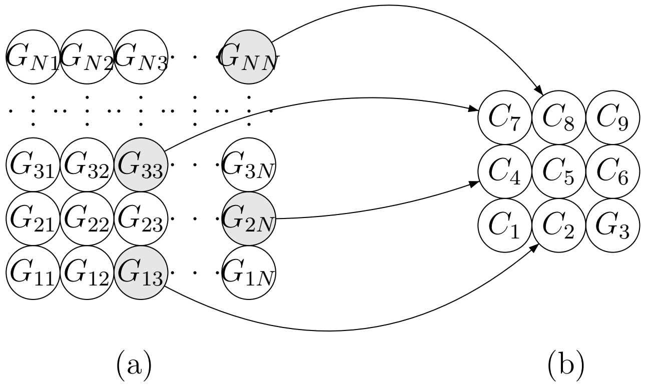


Fig. 6. Two Kohonen networks (from [50]) with a rectangular structure to model dialogue acts: The inputs to the large network (on the left) are a set of binary utterance features. Neurons representative of DA classes are grayed. The small network on the right represents the outputs of system (DA classes). The connexions between the neighboring nodes are not shown.

Kohonen networks are used for dialogue act recognition in [50]. The authors use seven *superficial* utterance features: speaker, sentence mode, presence or absence of a wh-word, presence or absence of a question mark, etc. Each utterance is represented by a pattern of these features, which is encoded into a binary format for the SOM representation. Initially, the exact number of DA classes is not known *a priori*, and only the large network on the left is created and trained. The clustering process is interrupted after a given number of clusters have been found.

To interpret the clusters, another small Kohonen network is built (the right model in Figure 6). This network contains as many neurons as DA classes. These neurons are initialized by the values of the weight-vectors of the representative neurons from the large network.

The quality of classification is evaluated by the Specificity Index (SI) [51] and by the Mean number of Conditions (MoC). They achieve about 0.1 for SI and about 2.6 for MoC on the SCHISMA corpus, with 15 DA classes and a network with 10 × 10 neurons. Another experiment has been performed with 16 DA classes and a larger network with 12 × 12 neurons with comparable results. Generally, unsupervised methods such as Kohonen networks are rarely used for DA recognition.

4.2 Decision Trees

Decision trees (or Classification and Regression Trees, CARTs) [52] are generation tools that are successfully used in operations research and decision analysis. They are usually represented by an oriented acyclic graph (see Figure 7). The root of the tree represents the starting point of the decision, each node contains a set of conditions to evaluate, and arcs show the possible outcomes of these decisions.

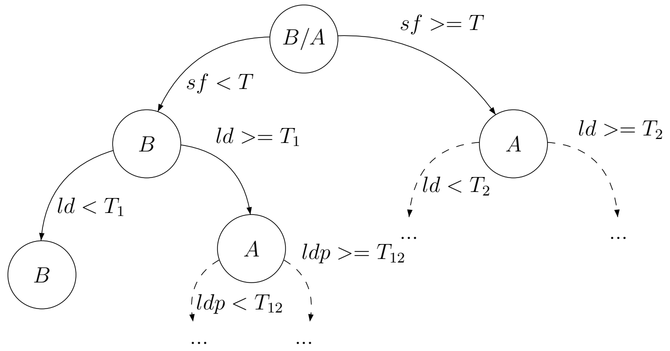


Fig. 7. Example of a part of the decision tree in the DA recognition domain: recognition of Backchannels (B) and Accepts (A) by prosody, from [16]

In the case of DA recognition, the decisions usually concern utterance features. Each decision compares the value of some feature with a threshold. For example, in Figure 7, three different prosodic features (sf, ld and ldp) are shown with their corresponding thresholds (T, T_1, T_2 and T_{12}). sf is the pause type feature and ld and ldp are the duration type features. Training of the decision tree is performed automatically on the training corpus. The output of the CART is the probability of the DA given the utterance features (lexical and prosodic), i.e., the *posterior* probability $P(C|W, F)$. The main advantage of CARTs is that they can combine different discrete and continuous features.

Wright uses in [5] 54 suprasegmental and duration prosodic features to train the trees on the CART algorithm [52]. She achieves 63 % of accuracy on the DCIEM map task corpus with 12 DA classes. Shriberg et al. also use in [16] CARTs for DA recognition with prosodic features. They use CARTs to recognize a few DAs only, which are very difficult to recognize with lexical (and syntactic) features. These DAs are recognized from prosody only. CARTs are used for example to distinguish statements from questions because questions usually differ from statements by an increasing final F0 curve. Therefore, this CART classifier is trained on statements and questions data only. Levin et al. compare in [30] CARTs with other classifiers, mainly Naive Bayes and MLP classifiers. They use binary grammatical features for this comparison. They show that CARTs outperform the Naive Bayes classifier and that they give comparable results with an MLP. The resulting DA recognition accuracy is about 68 % for English and about 66 % for German on the NESPOLE corpus.

4.3 Memory-Based Learning

Memory-Based Learning (MBL) [53] is an application of the memory-based reasoning theory in the field of machine learning. This theory is based on the assumption that it is possible to handle a new sample by matching it with stored representations of previous samples. Hence, in MBL, all known samples are stored in memory for future reference, and any unknown sample is classified by comparing it with all the stored samples. The main advantage of MBL compared to other machine learning techniques is that it successfully manages exceptions and sub-regularities in data. The main drawback of the method is its high memory and computational requirements.

Several methods can be used to compare the stored and recognized samples. The most popular one is the k-Nearest Neighbor (k-NN) [54]. It consists in defining a distance measure between samples, and of retrieving the k stored samples that have the smallest distance to the target one. These k samples are assumed to be similar to the recognized one, and the recognized sample is classified into the dominant class amongst these "neighbors".

Rotaru uses in [55] MBLs in an automatic dialogue acts tagging task on the Switchboard corpus [40] of spontaneous human-human telephone speech. The utterance features are based on word bigrams computed on the whole training corpus. These bigrams are hashed to a given number of features, whose optimal value is found experimentally. The hash function uses the letters present in the bigrams and the number of features. The author experiments a various number of neighbors. The best performance is about 72 % of accuracy with three neighbors. Levin et al. exploit in [30] MBLs on the NESPOLE corpus. They use the same features as described in the MLP case (see Section 4.1.1) on the IB1 algorithm [56] with one neighbor. They achieve about 70 % of accuracy for English and about 67 % for German. MBLs are also used in [57] with the IB1 algorithm. The authors obtain an accuracy of about 74 % with prosodic, lexical and context features on a corpus of Dutch telephone dialogues between users and the Dutch train timetable information system.

4.4 Transformation-Based Learning

The main idea of Transformation-Based Learning (TBL) [58] is to start from some simple solution to the problem, and to apply transformations to obtain the final result. Transformations are composed in a supervised way. Given a labeled training corpus and a set of possible transformation templates on this corpus, all possible transformations are generated from the templates, after what the transformations are selected iteratively. The templates can be for example: if tag X is after tag Y and/or N previous utterances contain word w, then change actual tag to Z. At each step the "best" transformation (bringing the largest improvement to precision) is selected and applied to the current solution. The algorithm stops when the selected transformation does not modify the data enough, or when there are no more transformations left.

The total number of all possible transformations can be very high. It is thus often computationally expensive to test all transformations, especially since most of them do not improve precision. A Monte-Carlo (MC) approach [59] can be used to tackle this issue: only a fixed number of transformations are selected randomly and used in the next steps. Although this may exclude the best transformation from the retained set, there are usually enough transformations left so that one of them still brings a large improvement to precision.

TBL can be applied to most classification tasks, and has been proposed for automatic DA recognition and some related works. [60] use TBL with a Monte Carlo strategy on the VERBMOBIL corpus. They use the following utterance features for DA recognition: cue phrases, word n-grams, speaker identity, punctuation marks, the preceding dialogue act, etc. The resulting DA accuracy is about 71 % with 18 dialogue acts. Bosch et al. use in [61] TBLs on the corpus of Dutch telephone dialogues between users and the Dutch train timetable information system, with a very limited DA tag-set. Question-answer pairs are represented by the following feature vectors: six features represent the history of questions asked by the system, while the following features represent the recognized user utterance, which is encoded as a sequence of bits, with 1 indicating that the i^{th} word of the lexicon occurs at least one time in the word graph. The last feature is used for each user utterance to indicate whether this sentence gave rise to a communication problem or not, as requested by the application, which final objective is to detect communication problems (incorrect system understanding) between the user and the dialogue system. They achieve to detect about 91 % of all communication problems with the rule-induction algorithm RIPPER [62]. The authors show that TBL outperforms MBL on this task. Lendvai et al. also use in [57] TBLs with the RIPPER algorithm. They obtain an accuracy of about 60 % with prosodic, lexical and context features on the same Dutch corpus as in the previous experiments.

4.5 Meta-Models

Model probabilities, such as the ones computed by the lexical n-gram previously described, can also be used as features of a "meta-model", whose role is to combine different sources of information in order to disambiguate the utterance. Hidden Markov Models are typically used for this purpose, as already described in Section 3.3. Another solution exploits boosting and committee-based sampling techniques, which can be used to compute tagging confidence measures, such as in [60], or to recognize sub-tasks labels [63], where a sub-task is defined as a sequence of DAs. Zimmermann compares in [64] n-gram, cue-phrases, maximum entropy and boosting classifiers for dialogue act recognition on a meeting corpus. On the ICSI MRDA meeting corpus, they obtain 23.3 % of DA recognition accuracy with 5 DA classes, by combining four individual DA classifiers: n-grams, cue phrases, maximum entropy and boosting. Combination is realized with an MLP.

5 DISCUSSION AND CONCLUSIONS

Automatic recognition of dialogue acts is an important yet still underestimated component of Human-Machine Interaction dialogue architectures. As shown in this review, research in this area has made great progress during the last years. Hence a few DA tag-sets have emerged as pseudo-standards and are more and more often used in the community. Nevertheless, these tag-sets are nearly always manually adapted to fit the specificities of each particular application, which points out a major issue in this area that concerns the variability of dialogue acts definitions and the consequent excessive costs and difficulty to port some previous work to a new task.

Another interesting characteristic of the dialogue act recognition domain is the fact that several different sources of information have to be combined to achieve reasonably good performances. In particular, most of the works discussed in this review show the importance of combining both lexical and prosodic information, as well as higher-level knowledge such as the overall structure of the dialogue used in the task, or semantic information such as named or task-related entities. This confirms our intuition that dialogue act recognition is a rich research area that might benefit from a better understanding of the dialogue processing, in particular with regard to the context of the dialog. Hence, many contextual relevant information are still not considered, for instance the social roles and relationships between the users, the emotions of the speakers, the surrounding environment as well as the past and recent history of interaction. All these information considerably influence the course of a dialogue, but are also extremely difficult to model and thus to include in our models. However, we have seen that the domain has progressively seen its influence area grows and intersects more and more with other research areas: from text to speech, from lexicon to prosody and semantic. We are convinced that this progression should continue, and that the overlap with adjacent domains should keep on enlarging, which is easier to achieve now thanks to the recent progress realized in, for example, the fields of user modeling and collaborative filtering, or emotions recognition, just to cite a few.

Acknowledgment

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

REFERENCES

[1] Austin, J. L.: How to Do Things With Words. Clarendon Press, Oxford 1962.

[2] Jeong, M.—Lee, G. G.: Jointly Predicting Dialog Act and Named Entity for Spoken Language Understanding. 2006.

[3] Garner, P. N.—Browning, S. R.—Moore, R. K.—Russel, R. J.: A Theory of Word Frequencies and its Application to Dialogue Move Recognition. In ICSLP '96, Vol. 3, pp. 1880–1883, Philadelphia, USA 1996.

[4] Alexandersson, J.—Reithinger, N.—Maier, E.: Insights into the Dialogue Processing of VERBMOBIL. Technical Report 191, Saarbrűcken, Germany 1997.

[5] Wright, H.: Automatic Utterance Type Detection Using Suprasegmental Features. In ICSLP '98, Vol. 4, p. 1403, Sydney, Australia 1998.

[6] Carletta, J.—Isard, A.—Isard, S.—Kowtko, J.—Newlands, A.—Doherty-Sneddon, G.—Anderson, A.: The Reliability of a Dialogue Structure Coding Scheme. Computational Linguistics, Vol. 23, 1997, pp. 13–31.

[7] Allen, J.—Core, M.: Draft of DAMSL: Dialog Act Markup in Several Layers. In `http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html`, 1997.

[8] Jurafsky, D.—Shriberg, E.—Biasca, D.: Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation (Coders Manual, Draft 13). Technical Report 97-01, University of Colorado, Institute of Cognitive Science, 1997.

[9] Dhillon, R.—Bhagat, S.—Carvey, H.—Shriberg, E.: Meeting Recorder Project: Dialog Act Labeling Guide. Technical Report TR-04-002, International Computer Science Institute, February 9, 2004.

[10] Jekat, S. et al.: Dialogue Acts in VERBMOBIL. In Verbmobil Report 65, 1995.

[11] Král, P.–Cerisara, C.—Klečková, J.: Lexical Structure for Dialogue Act Recognition. Journal of Multimedia (JMM), Vol. 2, No. 3, pp. 1–8, June 2007.

[12] Webb, N.—Hepple, M.—Wilks, Y.: Dialog Act Classification Based on Intra-Utterance Features. Technical Report CS-05-01, Dept. of Comp. Science, University of Sheffield 2005.

[13] Rosset, S.—Tribout, D.—Lamel, D.: Multi-Level Information and Automatic Dialog Act Detection in Human-Human Spoken Dialogs. Speech Communication 2008.

[14] Bangalore, S.—di Fabbrizio, S.—Stent, A.: Towards Learning to Converse: Structuring Task-Oriented Human-Human Dialogs. In ICASSP '06, Toulouse, France, May 2006.

[15] Stolcke, A. et al.: Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. In Computational Linguistics, Vol. 26, 2000, pp. 339–373.

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

[17] Langlais, P.: Traitement de la Prosodie en Reconnaissance Automatique de la Parole. Ph. D. thesis, Université d'Avignon et des pays de Vaucluse, 1995.

[18] Martin, P.: Prosodic and Rhythmic Structures in French. In Linguistics, Vol. II, 1987, pp. 925–949.

[19] Quang, V. M.—Besacier, L.—Castelli, E.: Automatic Question Detection Prosodiclexical Features and Crosslingual Experiments. In Interspeech 2007, pp. 2257–2260, Antwerp, Belgium, August 27–31, 2007.

[20] Mana, N. et al.: The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains. In Eurospeech 2003, Geneva, Switzerland, September 1–4, 2003.

[21] Meteer, M.—Taylor, A.—MacIntyre, R.—Iyer, R.: Dysfluency Annotation Stylebook for the Switchboard Corpus. Technical report, Linguistic Data Consortium, Ferbruary 1995. `ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps`, Revised June 1995 by A. Taylor.

[22] Stolcke, A.—Shriberg, E.: Automatic Linguistic Segmentation of Conversational Speech. In Proc. ICSLP '96, Vol. 2, pp. 1005–1008, Philadelphia, PA 1996.

[23] Shriberg, E.—Stolcke, A.—Hakkani-Tur, D.—Tur, G.: Prosody-Based Automatic Segmentation of Speech Into Sentences and Topics. In Speech communication, Vol. 32, pp. 127–154, September 2000.

[24] Kolar, J.—Liu, Y.—Shriberg, E.: Speaker Adaptation of Language Models for Automatic Dialog Act Segmentation of Meetings. In Interspeech 2007, pp. 1621–1624, Antwerp, Belgium, August 27–31, 2007.

[25] Janin, A.—Baron, D.—Edwards, J.—Ellis, D.—Gelbart, D.—Morgan, N.—Peskin, B.—Pfau, T.—Shriberg, E.—Stolcke, A.—Wooters, C.: The ICSI Meeting Corpus. In ICASSP 2003, pp. 364–367, Hong Kong, April 2003.

[26] Ang, J.—Liu, Y.—Shriberg, E.: Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. In Proc. ICASSP, March 2005.

[27] Liu, Y.: Structural Event Detection for Rich Transcription of Speech. Ph. D. thesis, Purdue University, December 2004.

[28] Berger, J. O.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York 1985.

[29] Reithinger, N.—Klesen, M.: Dialogue Act Classification Using Language Models. In EuroSpeech '97, pp. 2235–2238, Rhodes, Greece, September 1997.

[30] Levin, L.—Langley, C.—Lavie, A.—Gates, D.—Wallace, D.—Peterson, K.: Domain Specific Speech Acts for Spoken Language Translation. In 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan 2003.

[31] Grau, S.—Sanchis, E.—Castro, M. J.—Vilar, D.: Dialogue Act Classification using a Bayesian Approach. In 9th International Conference Speech and Computer (SPECOM 2004), pp. 495–499, Saint-Petersburg, Russia, September 2004.

[32] Ivanovic, E.: Dialogue Act Tagging for Instant Messaging Chat Sessions. In ACL Student Research Workshop, pp. 79–84, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

[33] Bilmes, J.—Kirchhoff, K.: Factored Language Models and Generalized Parallel Backoff. In Human Language Technology Conference, Edmonton, Canada 2003.

[34] Reithinger, N.—Maier, E.: Utilizing Statistical Dialogue Act Processing in VERBMOBIL. In 33rd annual meeting of the Association for Computational Linguistics, pp. 116–121, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

[35] Mast, M. 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.

[36] Stolcke, A. et al.: Dialog Act Modeling for Conversational Speech. In AAAI Spring Symp. on Appl. Machine Learning to Discourse Processing, pp. 98–105, 1998.

[37] Bard, E. G.—Sotillo, C.—Anderson, A. H.—Taylor, M. M.: The DCIEM Map Task Corpus: Spontaneous Dialogue Under Sleep Deprivation and Drug Treatment. In ICSLP '96, Vol. 3, pp. 1958–1961, Philadelphia, USA 1996.

[38] Ries, K.: HMM and Neural Network Based Speech Act Detection. In ICASSP '99, Vol. 3, 1999, pp. 497–500.

[39] Douglas, P. T.—Jay, F. N.: Speech Act Profiling: A Probabilistic Method for Analyzing Persistent Conversations and their Participants. In 37th Annual Hawaii International Conference on System Sciences (HICSS '04), IEEE 2004.

[40] Godfrey, J. J.–Holliman, E. C.—McDaniel, J.: SWITCHBOARD: Telephone Speech Corpus for Research and Development. In ICASSP 1992, Vol. 1, pp. 517–520, San Francisco, CA, USA, March 23–26, 1992.

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

[42] Keizer, S.: Dialogue Act Classification: Experiments with the SCHISMA Corpus. Technical report, University of Twente, October 2002.

[43] Ji, G.—Bilmes, J.: Dialog Act Tagging Using Graphical Models. In ICASSP '05, Vol. 1, pp. 33–36, Philadelphia, USA, March 2005.

[44] Dielmann, A.—Renals, S.: DBN Based Joint Dialogue Act Recognition of Multiparty Meetings. In ICASSP '07, pp. 133–136, Honolulu, Hawaii, USA, April 2007.

[45] Carletta, J. et al.: The AMI Meeting Corpus: A Preannouncement. In Multimodal Interaction and Related Machine Learning Algorithm Workshop (MLMI-05), Edinburgh, UK, July 11–13, 2005.

[46] Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall, 2nd edition, 1999.

[47] Sanchis, E.—Castro, M. J.: Dialogue Act Connectionist Detection in a Spoken Dialogue System. In Second International Conference on Hybrid Intelligent Systems (HIS 2002), pp. 644–651, Santiago de Chile, December 1–4, 2002. IOS Press.

[48] Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences, 30, 1995.

[49] Cottrell, M.—Fort, J. C.: Theoretical Aspects of the SOM Algorithm. Neurocomputing, Vol. 21, pp. 119–138, 1998.

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

[51] Andernach, T.: A Machine Learning Approach to the Classification of Dialogue Utterances. In NeMLaP-2, Ankara, Turkey, July 1996.

[52] Breiman, L.—Friedman, J.—Olshen, R.: Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove, CA 1984.

[53] Daelemans, W.—Zavrel, J.—van der Sloot, K.—van den Bosch, A.: TiMBL: Tilburg Memory-Based Learner. Technical report, Tilburg University, November 2003.

[54] Cover, T. M.—Hart, P. E.: Nearest Neighbor Pattern Classification. In IEEE Trans. Inform. Theory, 1967, pp. 21–27.

[55] Rotaru, M.:. Dialog Act Tagging Using Memory-Based Learning. Technical report, University of Pittsburgh, Spring 2002. Term Project in Dialog Systems.

[56] Aha, D. W.—Kibler, D.—Albert, K. M.: Instance-Based Learning Algorithms. Machine Learning, Vol. 6, No. 1, pp. 37–66, January 1991.

[57] Lendvai, P.—van den Bosch, A.—Krahmer, E.: Machine Learning for Shallow Interpretation of User Utterances in Spoken Dialogue Systems. In EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles Management, pp. 69–78, Budapest, Hungary 2003.

[58] Brill, E.: A Corpus-Based Approach to Language Learning. Ph. D. thesis, University of Pennsylvania, Philadelphia, USA 1993.

[59] Woller, J.: The Basics of Monte Carlo Simulations. University of Nebraska-Lincoln, Spring 1996. `http://www.chem.unl.edu/zeng/joy/mclab/mcintro.html`.

[60] Samuel, K.—Carberry, S.—Vijay-Shanker, K.: Dialogue Act Tagging with Transformation-Based Learning. In 17th international conference on Computational linguistics, Vol. 2, pp. 1150–1156, Montreal, Quebec, Canada, August 10–14, 1998. Association for Computational Linguistics, Morristown, NJ, USA.

[61] van den Bosch, A.—Krahmer, E.—Swerts, M.: Detecting Problematic Turns in Human-Machine Interactions: Rule-Induction Versus Memory-Based Learning Approaches. In 39th Meeting of the Association for Computational Linguistics, pp. 499–506, Toulouse, France 2001.

[62] Cohen, W.: Learning Trees and Rules with Set-valued Features. In 13th National Conference on Artificial Intelligence (AAAI-96), Vol. 1, pp. 709–716, Portland, Oregon 1996. AAAI Press.

[63] Tur, G.—Guz, U.—Hakkani-Tűr, D.: Model Adaptation for Dialog Act Tagging. 2006.

[64] Zimmermann, M.—Hakkani-Tűr, D.—Shriberg, E.—Stolcke, A.: Machine Learning for Multimodal interaction. Vol. 4299 of Lecture Notes in Computer Science, chapter Text Based Dialog Act Classification for Multiparty Meetings, pp. 190–199, Springer, Berlin/Heidelberg 2006.

Pavel Král has graduated from University of West Bohemia in the Department of Computer Science and Engineering in Plzeň (Czech Republic) and from Henri Poincaré University in Nancy (France) in 2007. He is now a lecturer, researcher at the University of West Bohemia. He is also a member of the Speech Group at LORIA-INRIA in Nancy. His research interests include speech recognition, more precisely automatic dialogue act recognition. He received his M. Sc. degree in 1999 with honours in the Department of Informatics and Computer Science at the University of West Bohemia.

Christophe Cerisara has graduated in computer science from the ENSIMAG engineering school in Grenoble in 1996, and obtained the Ph. D. at the Institut National Polytechnique de Lorraine in 1999. From 1999 to 2000 he worked as a researcher at Panasonic Speech Technology Laboratory in Santa Barbara. He is now a research scientist at CNRS, and belongs to the Speech Group in LORIA. His research interests include multi-band models and robust automatic speech recognition to noise. He is the author or co-author of more than forty scientific publications

Lexical Structure for Dialogue Act Recognition

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

[1]LORIA UMR 7503, BP 239 - 54506 Vandoeuvre, France
Email: {kral, cerisara}@loria.fr

[2]Dept. Informatics & Computer Science, University of West Bohemia, Plzeň, Czech Republic
Email: {pkral, kleckova}@kiv.zcu.cz

***Abstract*— This paper deals with automatic dialogue acts (DAs) recognition in Czech. Dialogue acts are sentence-level labels that represent different states of a dialogue, such as questions, hesitations, ... In our application, a multimodal reservation system, four dialogue acts are considered: statements, orders, yes/no questions and other questions. The main contribution of this work is to propose and compare several approaches that recognize dialogue acts based on three types of information: lexical information, prosody and word positions. These approaches are tested on a Czech Railways corpus that contains human-human dialogues, which are transcribed both manually and with an automatic speech recognizer for comparison. The experimental results confirm that every type of feature (lexical, prosodic and word positions) bring relevant and somewhat complementary information. The proposed methods that take into account word positions are especially interesting, as they bring global information about the structure of the sentence, at the opposite of traditional n-gram models that only capture local cues. When word sequences are estimated from a speech recognizer, the resulting decrease of accuracy of all proposed approaches is very small (about 3 %), which confirms the capability of the proposed approaches to perform well in real applications.**

***Index Terms*— dialogue act, language model, prosody, sentence structure, speech recognition**

I. Introduction

This work deals with automatic dialogue act recognition from the speech signal. A *dialogue act (DA)* represents the meaning of an utterance at the level of illocutionary force [1]. For example, "question" and "answer" are both possible dialogue acts. Automatically recognizing such dialogue acts is of crucial importance to interpret users' talk and guarantee natural human-computer interactions. For instance, this information might be used to check whether the user is requesting some information and is waiting for it, or to evaluate the feedback of the user. Another application is to animate a talking head that reproduces the speech of a speaker in real-time, by giving it facial expressions that are relevant to the current state of the discourse. In the following, a Czech train ticket reservation application has been used to assess the proposed methods.

As summarized in section II, two main types of features are generally used in the literature to automatically recognize dialogue acts: word sequences and prosody. A probabilistic dialogue grammar is also often used as additional stochastic information. Word sequences are most of the time modeled by statistical n-gram models, which encode the relationship between words and dialogue acts locally. In this work, we investigate a new kind of information for dialogue act recognition, that is the words position in the utterance. In contrast to n-grams, this information is global at the sentence level. Intuitively, this information is quite important for this task, as for instance, the word "who" is often at the beginning of sentences for questions, and at other positions for declarative sentences. A standard approach that takes into account this information consists in analyzing the sentence into a syntactic tree, but such analyzers are also known to work poorly in spontaneous speech. Hence, our approach is rather based on statistical methods.

We have already studied this problem in [2], [3], and proposed two approaches to solve it. In this work, we shortly present again both methods in section III, and further propose a third one that decouples the position from the lexical models, with the objective of optimizing the available training corpus. This paper also analyzes the gain obtained by merging lexical information with prosody, and discusses the combination of the proposed dialogue act recognition approach with a state-of-the-art speech recognizer, in order to deploy this system in realistic speech-driven applications. Section IV evaluates and compares these methods. In the last section, we discuss the research results and propose some future research directions.

II. Related Work

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

Different sets of dialogue acts are defined in these works, depending on the target application and the avail-

This paper is based on "Automatic Dialog Acts Recognition based on Sentence Structure," by P. Král, C. Cerisara, and J. Klečková, which appeared in the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, May 14-19, 2006. © 2006 IEEE.

able corpora. In [4], 42 dialogue acts classes are defined for English, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [5]. Switchboard-DAMSL tag-set [6] (SWBD-DAMSL) is an adaptation of DAMSL in the domain of telephone conversation. The Meeting Recorder DA (MRDA) tag-set [7] 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 [8] defines for German and Japanese 42 DAs, with 18 DAs at the illocutionary level, in the context of the VERBMOBIL corpus.

These general sets are usually further reduced into a much smaller number of broad classes, either because some classes occur rarely, or because the target application does not require such detailed classes. For instance, a typical regrouping may be the following [9]:

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

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

1) DA-models of the words sequences
2) dialogue grammars that model sequences of DAs
3) DA-models based on the utterance prosody

The first class of models infers the DA from the words sequences. These models are usually either probabilistic models, such as n-gram language models [4], [10], or knowledge-based approaches, such as semantic classification trees [10].

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* [11].

Semantic classification trees are decision trees that operate on words sequences with rule-based decision. These rules can be either trained automatically on a corpus, or manually coded.

A dialogue grammar is used to predict the most probable next dialogue act based on the previous ones. It can be modeled by Hidden Markov Models (HMMs) [4], Bayesian Networks [12], Discriminative Dynamic Bayesian Networks (DBNs) [13], or n-gram language models [14].

Prosodic models [9] can be 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 [15]:

- a falling intonation for most 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

In [9], the duration, pause, fundamental frequency (F0), energy and speaking rate prosodic attributes are modeled by a CART-style decision trees classifier. In [16], prosody is used to segment utterance. The duration, pause, F0-contour and energy features are used in [17], [18]. In both [17] and [18], 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 [9] 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.

Lexical and prosodic classifiers are combined in [4] as follows:

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

where C represents a dialogue act and W and F, which respectively represent lexical and prosodic information, are assumed independent.

III. Lexical position for dialogue act recognition

Syntax information is often modeled by probabilistic n-gram models. However, these n-grams usually model *local* sentence structure only. Syntax parsing could be used to associate sentence structures to particular dialogue acts, but conceiving general grammars is still an open issue, especially for spontaneous speech.

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.

A. *Sentence structure model*

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

We simplify this problem by assuming that each word is independent of 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 dialogue act class of the test sentence, w represents a word and p its position in the sentence.

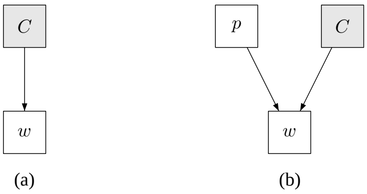


Figure 1. Graphical model of our approaches: grayed nodes are hidden

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 2.

$$P(w_1,\cdots,w_T|C) = \prod_{i=1}^{T} P(w_i|C) \qquad (2)$$

Dialogue act recognition then consists in finding the dialogue act $\hat{C}$ that maximizes the a posteriori probability:

$$\begin{aligned}\hat{C} &= \arg\max_C P(C|w_1,\cdots,w_T)\\ &= \arg\max_C P(C)\prod_{i=1}^{T} P(w_i|C) \qquad (3)\end{aligned}$$

This system is referred to as the “unigram” or “Naive Bayes” classifier [19].

On the right part of Figure 1, information about the position of each word is included. Then, the following issues have to be solved:

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

The first issue is solved by defining a fixed number of positions N_p: N_p likelihoods $P(w_i|C,p)$ are thus 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 is repeated at several positions.
- When $T > N_p$, several words can be aligned with one position. The likelihood at this position is the average over the N_i aligned words $(w_i)_{N_i}$:

$$P(w|C,p) = \frac{1}{N_i}\sum_{i}^{N_i} P(w_i|C,p) \qquad (4)$$

We propose and compare three methods to solve the second issue. The first *multiscale position* method considers the relative positions in a multiscale tree to smooth the models likelihoods. The second *non-linear merging* method models the dependency between W and p by a non-linear function that includes p. The third *best position* method decouples the positions from the lexical identities to maximize the available training corpus.

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: the model is equivalent to unigrams. Then, recursively, sentences are split into two parts of equal size and the number of possible positions is doubled.

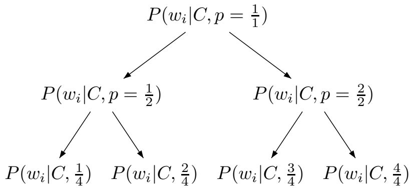


Figure 2. Multiscale position tree

For each word w_i, a threshold is applied on its number of occurrences and $P(w_i|C,p)$ for this word is computed at the finest scale that contains that minimum number of occurrences. This corresponds to the standard back-off technique [20] to solve the problem of lack of data.

Classification is then realized based on the following equation:

$$\begin{aligned}\hat{C} &= \arg\max_C P(C|w_1,\cdots,w_T,p_1,\cdots,p_T)\\ &= \arg\max_C P(C)\prod_{i=1}^{T} P(w_i|C,p_i) \qquad (5)\end{aligned}$$

where each likelihood is estimated at the finest scale possible.

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 $P(C|w_1,\cdots,w_T,p_1,\cdots,p_T)$ and the best class is simply given by:

$$\hat{C} = \arg\max_C P(C|w_1,\cdots,w_T,p_1,\cdots,p_T) \qquad (6)$$

B. Best position approach

We now give a slightly different definition for p: for any utterance W, let p be the best position amongst every possible position, i.e. the position that minimizes the DA recognition error rate.

Our objective is still to maximize:

$$\begin{aligned} P(C|W) &= \frac{P(W|C)P(C)}{P(W)} && (7) \\ &= \frac{P(C)\sum_p P(W,p|C)}{P(W)} && (8) \\ &= \frac{P(C)\sum_p P(W|C,p)P(p|C)}{P(W)} && (9) \end{aligned}$$

Now, once the best position p has been defined for a given utterance, the decision about the winning DA class can be taken based solely on this best position:

$$P(W|C,p) = P(w_p|C)$$

where w_p is the word of the current sentence at the best position p. Hence,

$$P(C|W) = \frac{P(C)\sum_p P(w_p|C)P(p|C)}{P(W)} \quad (10)$$

Finally, maximization gives:

$$\hat{C} = \arg\max_C P(C)\sum_p P(w_p|C)P(p|C) \quad (11)$$

In this equation, the lexical likelihood $\prod_i P(w_i|C)$ used so far is replaced by the weighted sum of each word likelihood. The weights intuitively represent the importance of each position, for a given DA class.

Compared to the previously proposed solutions that take into account the global position of the words, 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 tree.

Two factors might be considered to compute these weights: they can of course be trained on a labeled corpus, but we can also use some expert knowledge to define them. For instance, it is well-known that the words at the beginning of a sentence are important to recognize questions. This expert knowledge can be easily introduced as an *a priori* probability.

A posteriori weights can also be obtained after training on a development corpus. In the following experiments, the weights are trained based on the minimum DA error rate criterion, using a gradient-descent algorithm. The initial values of the weights are obtained by first evaluating on the development corpus the DA recognition accuracy when considering only the word at position p, for every possible p. The position p that gives the best recognition accuracy represents the most important position in the sentence. The gradient descent procedure then starts from this original position.

C. Prosody

Following the conclusions of previous studies [21], only the two most important prosodic attributes are considered: F0 and energy. The F0 curve is computed from 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 [22]. 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.

Two models are trained on these features and compared. The first one if a Muti-Layer Perceptron that outputs $P(C|F)$. The best class is then:

$$\hat{C} = \arg\max_C P(C|F) \quad (12)$$

The second one is a Gaussian Mixture Model (GMM) that models $P(F|C)$. The best class is then:

$$\hat{C} = \arg\max_C P(C|F) = \arg\max_C P(F|C)P(C) \quad (13)$$

D. Combination

The outputs of the lexical, position and prosodic models are normalized so that 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 [23].

IV. Evaluation

A. LASER speech recognizer

The LASER (LICS Automatic Speech Extraction/Recognition) software is currently under development by the Laboratory of Intelligent Communication Systems (LICS) at the 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 that combines the advantages of the hidden Markov model approach with those of artificial neural networks. A typical hybrid system uses HMMs with state emission probabilities computed from the output neuron activations of a neural network (such as the multi layer perceptron).

1) Neural network acoustic model: According to many authors (see e.g. [24]) 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, the cross entropy error criterion is used instead of the usual summed square error. Training this multi layer perceptron requires the precise knowledge of phoneme boundaries. 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, three-states HMMs phonetic units are modeled. However, all three states share the same emission probability computed from the activation value of one 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 changed to state emission probabilities:

$$P(o|S_j) = \frac{P(S_j|o) \cdot P(o)}{P(S_j)}. \tag{14}$$

where S_j denotes the j^{th} HMM state. The term $P(o)$ remains constant during the whole recognition process and hence can be ignored. The emission likelihoods are then computed 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.

2) Language model: Training words n-gram language models is not a good option in our case, because of the small size of our corpus, which is composed of manual transcriptions of a railway application (see Section IV-B). The chosen solution has been to merge words into classes and train an n-gram model based on those classes. This should compensate the lack of training data for infrequent word n-grams.

The method tries to automatically cluster words into classes according to their functional position in sentence. The algorithm (see [25]) begins with assigning each word into separate class and then starts merging two classes at a time. The process is stopped when the desired number of classes is reached. In the following experiments, the number of classes has been empirically set to 100 classes, and a trigram language model has been trained on these classes.

B. Dialogue acts corpus

The Czech Railways corpus contains human-human dialogues recorded in Czech, in the context of a train ticket reservation application. The number of sentences of this corpus is shown in column 2 of Table I.

The LASER recognizer is trained on 6234 sentences (c.f. first part of Table I), while 2173 sentences pronounced by different speakers (c.f. second part of Table I) are used for testing. Sentences of the test corpus have been manually labeled with the following dialogue acts: statements (S), orders (O), yes/no questions (Q[y/n]) and other questions (Q). The word transcription given by the LASER recognizer is used to compare the performances of DAs recognition experiments with the scores obtained from 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 of about ± 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 I.
COMPOSITION OF THE CZECH RAILWAYS CORPUS

C. Sentence structure experiments

1) Multiscale position: The multiscale position approach trains a model of $P(w_i|C,p)$ at different scales, as shown in Figure 2. Recognition is then performed based on equation 5.

Figure 3 shows the recognition accuracy of this method in function of the minimum number of word occurrence at each scale: this number defines the threshold used in the multiscale tree to select the finest possible scale to

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

estimate the observation likelihood. 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 for comparison.

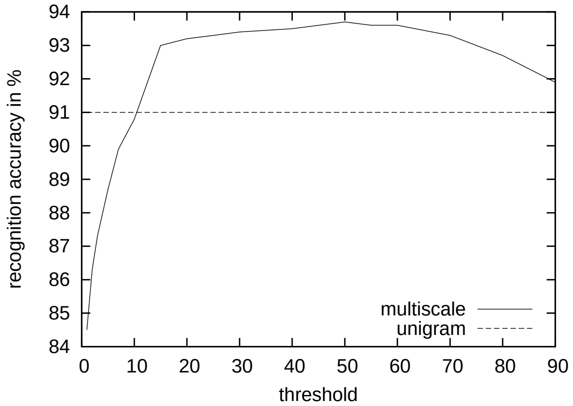


Figure 3. Dialogue acts recognition accuracy of the multiscale position system. The X-axis represents the minimum number of words in the tree, and the Y-axis plots the DA recognition accuracy

The recognition accuracy of each class is shown in Table II.

These experimental results confirm that taking into account the global position of each word improves the recognition accuracy. Furthermore, the proposed multiscale tree seems to be a reasonable solution to the issue concerning the lack of training data.

2) Non-linear merging: In the second experiment, the *Non-linear* model that merges lexical and position information is implemented by a Multi-Layer Perceptron (MLP). The chosen MLP topology is composed of three layers: 4 (for each DA class) times 8 (equal-size segments of the sentence) input neurons, 12 neurons in the hidden layer and 4 output neurons, which encode the *a posteriori* class probability. The dialogue act class is given by equation 6.

The recognition results of these methods are shown in Table II, along with the results obtained with the baseline unigram model.

3) Best position approach: The third position-based approach proposed is the *best-position* method, which recognizes dialogue acts based on equation 11. In this method, the number of positions allowed is not limited by the size of the training corpus. Hence, twenty positions (instead of eight positions for the two previous approaches) are considered.

In order to compute the initial values of the weights $P(p|C)$, recognition is first performed on the development corpus using only one position at a time:

$$P(p = i|C) = 1 \text{ and } P(p \neq i|C) = 0 \text{ for all } C$$

where i is one of the twenty possible positions. This experiment is repeated for every possible i, and the recognition accuracies obtained with each i are shown in Figure 4.

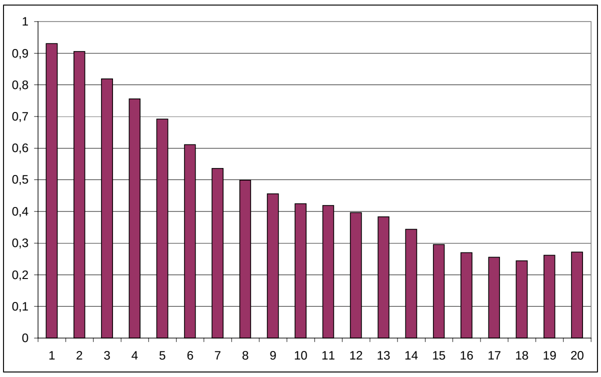


Figure 4. DA recognition accuracy on the development corpus when only a single position is considered.

Based on this experiment, the initial values chosen for the gradient descent algorithm are:

$$P(p = 1|C) = 1 \text{ and } P(p > 1|C) = 0 \text{ for all } C$$

After the gradient descent algorithm, the resulting weights are shown in Figure 5.

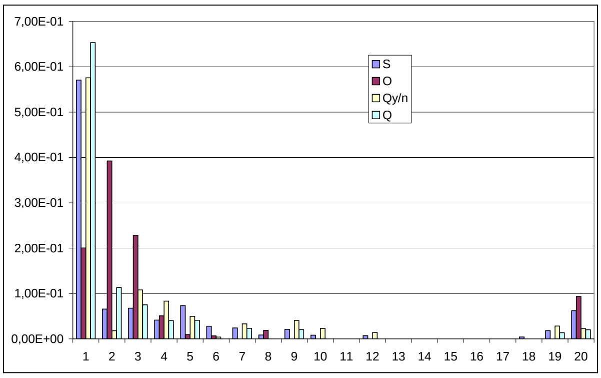


Figure 5. Weights obtained after the gradient-descent algorithm.

In this figure, it is clear that the most important positions for all DA classes are close to the beginning of the utterance. The last words of the utterance also have some importance, especially for the “order” class. The very first position is the most important for questions. These results are conforming to our intuition.

Then, using the weights shown in Figure 5, recognition is performed on the test corpus. The results are given in the fifth section of Table II.

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

D. Prosody

The third section of Table II shows the recognition accuracy obtained when only a prosodic model is used to classify dialogue acts. Two prosodic models are compared: the GMM (equation 13) and the MLP (equation 12).

The best MLP topology uses three layers: 40 inputs, 18 neurons in hidden layer and 4 outputs. The best results are obtained with the 3-mixtures GMM. It is difficult to use more Gaussians, because of the lack of training data, mainly for class O.

Although these recognition scores are much lower than the ones obtained with lexical features, it is shown next that prosody may nevertheless bring some relevant clues that are not captured by lexical models.

E. Combination

The fourth part of Table II shows the recognition accuracy when the prosodic GMM and the MLP-position models are combined with another MLP (as described in [23]).

One can conclude without loss of generality that the combination of models gives better recognition accuracy than both the lexical and prosodic models 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					
1 Unigram	93.5	77.6	96.5	89.9	**91.0**
2. Sentence structure					
2.1 Multiscale	94.7	70.4	96.1	95.3	**93.8**
2.2 Non-linear	90.3	83.2	91.1	98.8	**94.7**
3. Prosodic information					
3.1 GMM	47.7	43.2	40.8	44.3	**44.7**
3.2 MLP	38.7	49.6	52.6	34.0	**43.5**
4. Combination of 2.2 and 3.1					
MLP	91.5	85.6	94.0	98.7	**95.7**
5. Best position approach					
Best position	93.6	95.2	97.2	94.3	**95.8**

TABLE II.
DIALOGUE ACTS RECOGNITION ACCURACY FOR DIFFERENT APPROACHES/CLASSIFIERS AND THEIR COMBINATION WITH MANUAL WORD TRANSCRIPTION

F. Recognition with LASER recognizer

Table III 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 III structure is the same as Table II.

The errors in transcriptions induced by the automatic speech recognizer do not have a strong impact on the results presented so far: the final accuracy only decreases from 95.7 % down to 93 %, and the ordering of the methods' accuracy is preserved. This validates the use of the proposed approaches in human-computer speech-based applications that use such a speech recognizer.

	accuracy in [%]				
Approach/ Classifier	S	O	Q[y/n]	Q	Global
1. Lexical information					
1 Unigram	93.1	68.8	94.7	86.3	**88.2**
2. Sentence structure					
2.1 Multiscale	93.8	63.2	92.9	92.9	**91.4**
2.2 Non-linear	85.5	72.0	86.8	98.0	**91.8**
3. Prosodic information					
3.1 GMM	47.7	43.2	40.8	44.3	**44.7**
3.2 MLP	38.7	49.6	52.6	34.0	**43.5**
4. Combination of 2.2 and 3.1					
MLP	88.5	77.6	90.4	97.3	**93.0**
5. Best position approach					
Best position	92.1	86.4	95.3	92.2	**93.6**

TABLE III.
DIALOGUE ACTS RECOGNITION ACCURACY FOR DIFFERENT APPROACHES/CLASSIFIERS AND THEIR COMBINATION WITH WORD TRANSCRIPTION FROM LASER RECOGNIZER

V. CONCLUSIONS

In this work, we studied the influence of word positions in a dialogue act recognition task. Two previously proposed approaches and a third new one have been described and compared, both in terms of their respective theoretical advantages and drawbacks, and also experimentally on a Czech corpus for a train ticket reservation. It has thus been demonstrated that the global position of the words in sentences is an important information that improves automatic dialogue act recognition accuracy, at least when the size of the training corpus is too limited to train lexical n-gram models with a large n, which is the most common situation in dialogue act recognition.

One of the systems that combines both lexical and position information has then been enhanced by further considering prosodic information. Yet, several prosodic models have been compared, and the combined approach still improves the results over the position and lexicon approach alone.

Finally, the manual transcription has been replaced by an automatic transcription obtained from a Czech speech recognizer, in order to validate the use of the proposed dialogue act recognition approach in realistic applications that are often based on automatic speech recognition. The resulting decrease in performances is very small, which confirms the validity of the proposed approaches.

The focus of this work has been on modeling global words position, but local statistical grammars have not been largely exploited, mainly because of the lack of training data. However, these grammars shall also bring relevant information, and it would be quite advantageous to further combine the proposed global model with such local grammars. Another important information that has not been taken into account in this work is a dialogue act grammar, which models the most probable sequences of dialogue acts. It is straightforward to use such a statistical grammar with our system, but we have not yet done so because it somehow masks the influence of the statistical and prosodic features we focus on in this work, and also in order to keep the approach as general as possible. Indeed, such a grammar certainly improves the recognition results but is also often dependent on the target application. We also plan to test these methods on another corpus (broadcast news), another language (French) and with more DA classes.

ACKNOWLEDGMENT

This work has been partly supported by the European integrated project Amigo (IST-004182), a project partly funded by the European Commission, and by the Ministry of Education, Youth and Sports of Czech republic grant (NPV II-2C06009).

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

[5] 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.

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

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

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

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

[10] 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.

[11] 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.

[12] S. Keizer, A. 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.

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

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

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

[16] 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.

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

[18] 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.

[19] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, "Dialogue Act Classification using a Bayesian Approach," in *9th International Conference Speech and Computer (SPECOM'2004)*, Saint-Petersburg, Russia, September 2004, pp. 495–499.

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

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

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

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

[24] 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.

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

Pavel Král is a Ph.D. candidate from University of West Bohemia at Dept. Informatics & Computer Science in Plzeň (Czech Republic) and from Henri Poincaré University in Nancy (France). He is also a lecturer at the University of West Bohemia and a member of the Speech Group at LORIA-INRIA in Nancy. His research domain is on speech recognition, more precisely on automatic dialog acts recognition.

He received his M.Sc. degree in 1999 with honours in Dept. Informatics & Computer Science at the University of West Bohemia.

Christophe Cerisara is graduated from the engineering school ENSIMAG in computer science in Grenoble in 1996, and obtained the Ph.D. at the Institut National Polytechnique de Lorraine in 1999. He worked as a researcher from 1999 to 2000 at Panasonic Speech Technology Laboratory in Santa Barbara. He is now a research scientist at CNRS, and belongs to the Speech Group in LORIA. His research interests include multi-band models and robust automatic speech recognition to noise. He is the author or co-author of more than forty scientific publications.

Associated Professor **Jana Klečková** is a member of Department of Computer Science and Engineering, Faculty of Applied Sciences at the University of West Bohemia in Pilsen, Czech Republic. Her research fields are database systems, computational neuroscience (binding problem), speech recognition and understanding. She received her M.Sc. degree in 1972 at Electro technical faculty of VSSE Pilsen and her Ph.D. in 1997 at Faculty of applied sciences, UWB Pilsen.

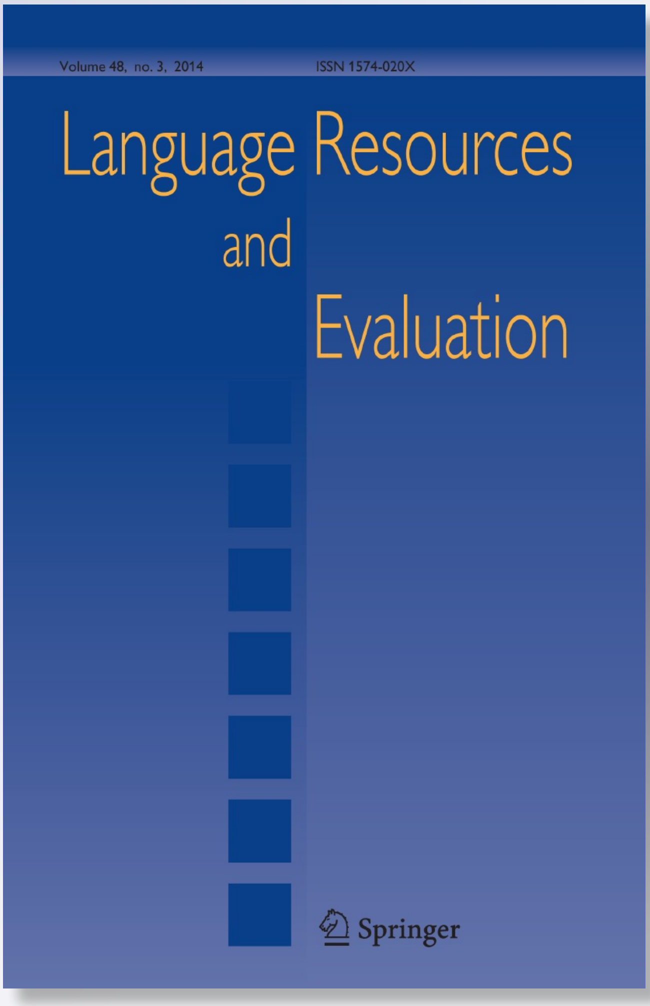
Automatic dialogue act recognition with syntactic features

Pavel Král & Christophe Cerisara

Language Resources and Evaluation

ISSN 1574-020X
Volume 48
Number 3

Lang Resources & Evaluation (2014)
48:419-441
DOI 10.1007/s10579-014-9263-6



Springer

Lang Resources & Evaluation (2014) 48:419–441
DOI 10.1007/s10579-014-9263-6

ORIGINAL PAPER

Automatic dialogue act recognition with syntactic features

Pavel Král · Christophe Cerisara

Published online: 8 February 2014

Abstract This work studies the usefulness of syntactic information in the context of automatic dialogue act recognition in Czech. Several pieces of evidence are presented in this work that support our claim that syntax might bring valuable information for dialogue act recognition. In particular, a parallel is drawn with the related domain of automatic punctuation generation and a set of syntactic features derived from a deep parse tree is further proposed and successfully used in a Czech dialogue act recognition system based on conditional random fields. We finally discuss the possible reasons why so few works have exploited this type of information before and propose future research directions to further progress in this area.

Keywords Dialogue act · Language model · Sentence structure · Speech act · Speech recognition · Syntax

P. Král (✉)
Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia, Plzeň, Czech Republic
e-mail: pkral@kiv.zcu.cz

P. Král
Faculty of Applied Sciences, New Technologies for the Information Society (NTIS),
University of West Bohemia, Plzeň, Czech Republic

C. Cerisara
LORIA UMR 7503, BP 239, 54506 Vandoeuvre, France
e-mail: cerisara@loria.fr

1 Introduction

1.1 Definition

Modelling and automatically identifying the structure of spontaneous dialogues is very important to better interpret and understand them. The precise modelling of dialogues is still an open issue, but several specific characteristics of dialogues have already been clearly identified. *Dialogue Acts* (*DAs*) are one of these characteristics.

Although the term "dialogue acts" that is commonly used nowadays has been defined by Austin (1962), a number of other seminal works have proposed very similar notions, including speech acts proposed by Searle (1969), conversational game moves introduced by Power (1979), adjacency pairs proposed by Schegloff (1968, Sacks et al. 1974) or acts of communication in the plan-based approaches to understanding introduced by Litman et al. (1985), Kautz (1987), Carberry (1990). The theory of the dialogue acts has been further developed by Bunt (1994). The dialogue acts represent the meaning of an utterance in the context of a dialogue, where the context is divided into several types, with both global and local views: linguistic, semantic, physical, social and cognitive. Bunt also developed a multidimensional taxonomy of the dialogue acts, while David R. Traum developed the notion of speech acts in Traum (1999) with dialogue agents. A better overview of the notion of dialogue acts can be found in Stolcke (2000).

In this work, the dialogue act is seen as a function of an utterance, 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.

Table 1 illustrates the dialogue acts that may occur in a dialogue between the passenger (P) and the agent (A) in a ticket reservation task. The corresponding dialogue act labels are also shown. Each utterance is labelled with a unique dialogue act. This example is taken from our Czech corpus (see Sect. 5.1).

Dialogue acts represent useful and relevant information for many applications, such as dialogue systems, machine translation, automatic speech recognition, topic tracking (Garner et al. 1996) or talking head animation. For instance, in dialogue systems, dialogue acts might be used to recognize the intention of the user and thus differentiate situations where the user is requesting some information from situations where the user is simply giving some information or backchannels. In

Table 1 Example of a dialogue between the passenger (P) and the agent (A) in a ticket reservation task with the English translation

Speaker	DA	Dialogue in Czech	English translation
P	Question	Kdy pojede první vlak do Prahy?	When will the first train go to Prague?
A	Question yes/no	Chcete rychlík?	Do you want the express train?
P	Statement	To je jedno	I don't care
A	Statement	V osm hodin	At eight o'clock
P	Order	Dejte mi tedy jeden lístek, prosím	Give me one ticket, please
A	Statement	Tady je	Here it is

the former case, the system has to react, while in the latter case, a system reaction may be perceived as intrusive. In the machine translation domain, recognizing dialogue acts may bring relevant cues to choose between alternative translations, as the adequate syntactic structure may depend on the user intention. Automatic recognition of dialogue acts may also be used to improve the word recognition accuracy of automatic speech recognition systems, as proposed for instance in Wright (1998), where a different language model is applied during recognition depending on the dialogue act. Finally, dialogue act recognition is a fundamental building block of any understanding system and typically completes semantic role labelling and semantic frame inference.

The usefulness of dialogue act recognition has thus been demonstrated in a number of large applicative systems, such as the *VERBMOBIL* (Alexandersson et al. 1997), *NESPOLE* (Lavie et al. 2006) and *C-STAR* (Blanchon and Boitet 2000) machine translation and dialogue systems that rely on dialogue act classification.

1.2 Objectives

The main objective of this work is to propose and investigate the usefulness of syntactic features to improve dialogue act recognition in Czech. In previous works, we have first designed a baseline dialogue act recognition system for the Czech language that was based on generative models (Král et al. 2005). Although reasonably good results have been obtained, this approach was limited because it only exploits the local context around any given word of the utterance. We then proposed in Král et al. (2006, 2007) several approaches to address this limitation and include global features in the model that represent the sentence structure. One of these approaches consists in modelling the word position in the sentence as a random variable and integrating this variable in the generative model. Intuitively, this information is important for dialogue act recognition, as for instance, the word "who" is often located at the beginning of sentences for questions and at other positions for declarative sentences. In the following, we propose a different approach to model such global information implicitly, via a conditional stochastic model. The second and most important contribution of this work concerns the design and exploitation of syntactic features for dialogue act recognition in Czech. As summarized in Sect. 2, only a few types of features are generally used in the literature to automatically recognize dialogue acts: lexical, part-of-speech (POS) tags, dialogue history and prosody. Furthermore, word sequences are most of the time modelled by statistical n-gram models, which encode the relationship between words and dialogue acts only locally. While we have already shown the importance of global information such as word position in the utterance for dialogue act recognition, the current work goes beyond this type of information by investigating whether the conditional distribution of the target dialogue act depends on the syntactic structure of the utterance.

In the following section, we briefly review the state of the art about dialogue act recognition, with a focus on how syntactic information has already been considered for this task and for related tasks. In Sect. 3, we propose and describe new syntactic features. The proposed model is described in Sect. 4. The relative importance of each of these features is evaluated on a Czech dialogue corpus in Sect. 5 In the last section, we discuss these results and propose some future research directions.

2 Related work

We will now briefly review the standard definitions of dialogue acts, the different types of models classically used for dialogue act recognition and the standard types of information used in such models. Then, we review and discuss the previous design of syntactic features for dialogue act recognition as well as in closely related domains.

Some generic sets of domain-independent dialogue acts have been proposed in the state-of-the-art and are now commonly used to create the baseline tag set for most types of applications. Hence, in Stolcke (2000), 42 DAs classes are defined for English, based on the discourse annotation and markup system of labelling (DAMSL) tag-set (Allen and Core 1997). The switchboard–DAMSL tag-set (Jurafsky et al. 1997) (SWBD–DAMSL) is an adaptation of DAMSL in the field of telephone conversations. The meeting recorder dialogue act (MRDA) tag-set (Dhillon and Carvey 2004) is another very popular tag-set, which is based on the SWBD–DAMSL taxonomy. MRDA contains 11 general dialogue act labels and 39 specific labels. Finally, Jekat (1995) defines for German and Japanese 42 dialogue acts, with 18 dialogue acts at the illocutionary level, in the context of the VERBMOBIL corpus. The ISO standard 24617-2 for dialogue annotation has been published in 2012. DIT++[1] is a recent implementation of this standard. Because of the limited size of the available corpus, as well as several other technical reasons, these tag sets are frequently reduced by merging several tags together, so that the number of final actual generic tags is often about 10. Part of such typical generic dialogue acts, also referred to as speech acts, include for instance (Shriberg et al. 1998) statements, questions, backchannels, commands, agreements, appreciations as well as a broad "miscellaneous" class. In addition to such generic tags, application-specific tags may be defined, such as "request booking" for a hotel booking application.

Manually annotating dialogue acts on every new corpus may be very costly and efforts have been put into developing semi-automatic methods for dialogue act tagging and discovery. Hence, the authors of Orkin and Roy (2010) propose a predictive paradigm where dialogue act models are first trained on a small-size corpus and used afterwards to predict future sentences or dialogue acts. In a related vein, unsupervised dialogue act tagging of unlabelled text has recently raised a lot of attention (Joty et al. 2011; Crook et al. 2009), but we will limit ourselves in the following on supervised approaches.

The dialogue act recognition task is often considered jointly with the segmentation task. We assume in our work that sentence segmentation is known, because we rather prefer to concentrate on the challenge of designing relevant

[1] http://dit.uvt.nl.

syntactic features for dialogue act recognition. Yet, many related works propose powerful solutions for the segmentation task as well. In particular, the work described in Petukhova and Bunt (2011) considers the input text as a stream of words and segments and tags it incrementally with a BayesNet model with lexical, prosodic, timing and dialogue act-history features. Zimmermann et al. (2006) successfully use in for joint DA segmentation and classification hidden-event language models and a maximum entropy classifier. They use word sequence and pause duration as features. The authors of Dielmann and Renals (2008) exploit a Switching Dynamic Bayesian Network for segmentation, cascaded with a conditional random field for dialogue act classification, while Quarteroni et al. (2011) jointly segments and tags with a single model.

The dialogue act modelling schemes that are commonly used for dialogue act recognition are traditionally chosen from the same set of general machine learning methods used in most natural language processing tasks. These include Hidden Markov Models (Stolcke 2000), Bayesian Networks (Keizer and Nijholt 2002), Discriminative Dynamic Bayesian Networks (Ji and Bilmes 2005), BayesNet (Petukhova and Bunt 2011), memory-based (Lendvai and van den Bosch 2003) and transformation/based learning (Samuel et al. 1998), decision trees (Mast 1996), neural networks (Levin et al. 2003), but also more advanced approaches such as boosting (Tur et al. 2006), latent semantic analysis (Serafin and Di Eugenio 2004), hidden backoff models (Bilmes 2005), maximum entropy models (Ang et al. 2005), conditional random fields (CRFs) (Dielmann and Renals 2008; Quarteroni et al. 2011) and triangular-chain CRF (Jeong and Lee 2008).

Regarding features, most dialogue act recognition systems exploit both prosodic and lexical features. The dialogue history is also often used as relevant information. Some cue words and phrases can also serve as explicit indicators of dialogue structure (Webb 2010). For example, 88.4 % of the trigrams "⟨start⟩ do you" occur in English in *yes/no questions* (Jurafsky 1997).

Prosody is an important source of information for dialogue act recognition (Shriberg et al. 1998). For instance, prosodic models may help to capture the following typical features of some dialogue acts (Kompe 1997):

- a falling intonation for most statements
- a rising F0 contour for some questions (particularly for declaratives and yes/no questions)
- a continuation-rising F0 contour characterizes (prosodic) clause boundaries, which is different from the end of utterance

In Shriberg et al. (1998), the duration, pause, fundamental frequency (F0), energy and speaking rate prosodic attributes are modelled by a CART-style decision trees classifier. In Mast et al. (1996), prosody is used to segment utterances. The duration, pause, F0-contour and energy features are used in Wright (1998) and Wright et al. (1999). In both Wright (1998) and Wright et al. (1999), 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 dialogue act recognition. Shriberg et al. (1998) show that it is better to use prosody for dialogue act recognition in three separate tasks, namely question detection, incomplete utterance detection and agreements detection, rather than for detecting all dialogue acts in one task.

Apart from prosodic and contextual lexical features, only a few works actually exploit syntactic relationships between words for dialogue act recognition. Some syntactic relations are captured by HMM word models, such as the widely-used n-grams (Stolcke 2000), but these approaches only capture local syntactic relations, while we consider next global syntactic trees. Most other works thus focus on morphosyntactic tags, as demonstrated for instance in Verbree et al. (2006), where a smart compression technique for feature selection is introduced. The authors use a rich feature set with POS-tags included and obtain with a decision tree classifier an accuracy of 89.27, 65.68 and 59.76 % respectively on the ICSI, Switchboard and on a selection of the AMI corpus. But while POS-tags are indeed related to syntax, they do not encode actual syntactic relations.

A very few number of works have nevertheless proposed some specific structured syntactic features, such as for instance the subject of verb type (Andernach 1996). The authors of Serafin and Di Eugenio (2004), Di Eugenio et al. (2010) exploit a few global syntactic features, in particular POS-tags and the MapTask SRule annotation that indicates the main structure of the utterance, i.e., declarative, imperative, inverted or Wh-question, but without obtaining a clear gain from syntax in their context, hence suggesting that further investigation is needed. Indeed, syntax is a very rich source of information and the potential impact of syntactic information highly depends on the chosen integration approach and experimental setup. We thus propose in the next section other types of syntactic features and a different model and show that syntax might indeed prove useful for dialogue act recognition in the proposed context. But let us first support our hypothesis by briefly reviewing a few other papers that also support the use of syntax for both dialogue act recognition and closely related domains.

First, as already shown, word n-grams features, with n > 1, do implicitly encode local syntactic relations and are used successfully in most dialogue act recognition systems. But more importantly, a recent work (Klüwer et al. 1944) concludes that both dialogue context and syntactic features dramatically improve dialogue act recognition, compared to words only, more precisely from an accuracy of 48.1 % up to 61.9 % when including context and 67.4 % when further including syntactic features. They use in their experiments a Bayesian Network model and their syntactic features are the syntactic class of the predicate, the list of arguments and the presence of a negation. Although this work actually focuses on predicate-argument structures, while our main objective is rather to exploit the full syntactic tree without taking into account any semantic-level information for now, this work supports our claim that syntactic information may prove important for dialogue act recognition. In addition, Zhou et al. (2009) employ in three levels of features: (1) word level (unigram, bigram and trigram), (2) syntax level [POS-tags and chunks recognized as base noun phrase (BNP)] and (3) restraint information (word position, utterance length, etc.). Syntactic and semantic relations are acquired by information extraction methods. They obtain 88 % of accuracy with a SVM classifier on a Chinese corpus and 65 % on the SWBD corpus.

We further investigated closely related domains that have already explored this research track in more depth. This is for instance the case of automatic classification of rhetorical relations, as reviewed in Sporleder and Lascarides (2008). Another very close task is punctuation recovery, which aims at generating punctuation marks in raw words sequences, as typically obtained from speech recognition systems. In particular, this implies to discriminate between questions (ending with a question mark), orders (ending with an exclamation points) and statements (ending with a period), which is a task that is obviously strongly correlated to dialogue act recognition. Interestingly enough, a richer set of syntactic features have been exploited in the punctuation recovery domain than in the dialogue act recognition area. Hence, the authors of Favre et al. (2009) design several syntactic features derived from the phrase structure trees and show that these features significantly reduce the detection errors. This is in line with our own previous conclusions published in Cerisara et al. (2011) regarding the use of syntactic features for punctuation recovery, where a large improvement in performances is obtained thanks to syntactic information derived from dependency trees. Similar gains are obtained on a Chinese punctuation task (Guo et al. 2010), where including rich syntactic features, such as the word grammatical function, its ancestors and children, its head, the yield of the constituent or subtree border indicators, improve the F-measure from 52.61 % up to 74.04 %.

Finally, we have shown that there is an increasing amount of work that successfully exploits structural syntactic dependencies both for dialogue act recognition and in related domains such as punctuation recovery. We further believe that parsing of natural language utterances will constitute a fundamental pre-processing step of most if not all subsequent NLP modules, although it has probably not been as widely used as POS tagging for instance because of its complexity and lack of robustness to ill-formed input. However, thanks to the current progress in Bayesian approaches and feature-rich log-linear models, we expect parsing to be more and more robust to automatic speech recognition errors in the near future. Other recent reviews of the literature about dialogue act recognition are realized in Geertzen (2009) and Webb (2010).

3 Syntax for automatic dialogue act recognition

In our previous work (Král et al. 2006a, b, 2007), we proposed to include in our DA recognition approach information related to the position of the words within the sentence. In this work, we propose a different approach that derives features from the sentence parse tree and includes these features as input to a conditional random field. The parse trees are defined within the dependency framework (Hajičová 2000) and are automatically computed on the input sentences. We evaluate this approach in two conditions, respectively when the input sentences are manually and automatically transcribed.

3.1 Features

We distinguish next two types of features, respectively the baseline and syntactic features. The baseline features are:

- words inflected form
- lemmas
- part-of-speech tags
- pronoun or adverb at the beginning of the sentence
- verb at the beginning of the sentence

The syntactic features rely on a dependency parse tree:

- dependency label
- root position in the utterance
- unexpressed subjects
- basic composite pair (subject–verb inversion)

All these features are described in details next.

3.1.1 Baseline features

Words inflected form The word form is used as a baseline lexical feature in most modern lexicalized natural language processing approaches (Stolcke 2000; Keizer and Nijholt 2002; Ji and Bilmes 2005; Jurafsky 1997). In our case, sentence segmentation is known but capitalization of the first word of the sentence is removed, which decreases the total number of features in our model without impacting accuracy, thanks to the insertion of a special "start-of-utterance" word. Although word bigrams or trigrams are commonly used in other systems, we only use word unigrams because of the limited size of the training corpus. We rather compensate for this lack of local structural information by investigating global syntactic dependencies. The word forms are obtained in our experiments using both manual and automatic transcriptions of speech audio files.

Lemmas We used the *lemma* structure from the Prague Dependency Treebank (PDT) 2.0[2] (Hajič et al. 2000) project, which is composed of two parts. The first part is a unique identifier of the lexical item. Usually it is the base form (e.g., infinitive for a verb) of the word, possibly followed by a digit to disambiguate different lemmas with the same base forms. The second optional part contains additional information about the lemma, such as semantic or derivational information. Lemmas may in some circumstances bring additional information, notably by removing irrelevant variability introduced by inflected forms. This may have some importance in particular for rare words that may occur with different inflected forms but still may have some impact on the dialogue act decision process. The lemmas are obtained automatically in our experiment with a lemmatizer.

Part-of-speech (POS) tags The part-of-speech is a word linguistic category (or more precisely lexical item), which can be defined by the syntactic or morphological behaviour of the lexical item in question. There are ten POS categories defined in the

[2] http://ufal.mff.cuni.cz/pdt2.0/.

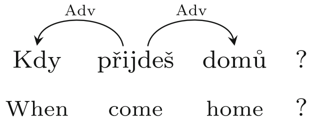


Fig. 1 Example of adverb as first word

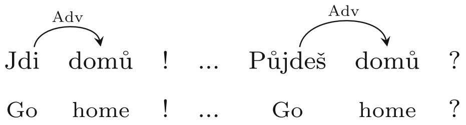


Fig. 2 Example of verb as first word

PDT (Hajič et al. 2000) for the Czech language: nouns, adjectives, pronouns, numerals, verbs, adverbs, prepositions, conjunctions, particles and interjections. The part-of-speech tags are inferred automatically in our experiment with a POS tagger.

Pronoun or adverb at the beginning of the sentence This boolean feature indicates whether the utterance starts with a pronoun or an adverb. It can be particularly useful for detecting *Wh-questions*, which usually start with a *Pronoun* (POS-tag "P") or an *Adverb* (POS-tag "D"), such as in: "Kdy přijdeš domů?" (When do you come home?) (Fig. 1).

Note that similar features that emphasize the importance of initial words in the sentence have already been proposed, for instance in Ang et al. (2005), Webb (2010), Jurafsky and Martin (2009).

Verb at the beginning of the sentence This feature is also a boolean indicator of the presence of a verb as the first word of an utterance. It can be particularly useful for the detection of *Commands* and *Yes–no questions*, which usually start with a verb, such as in: "Jdi domů!" (Go home!) and "Půjdeš domů?" (Do you go home?) (Fig. 2).

3.1.2 Syntactic features

All syntactic features are computed from the syntactic tree obtained after automatic parsing of the target sentence: a detailed description of our automatic parser is given in Sect. 5.2. We have chosen to represent the syntactic relations with *dependencies*, as it is commonly done nowadays for many natural language processing tasks. Furthermore, we have chosen the PDT to train our stochastic parser and our annotation formalism thus follows the one used in the PDT.

An example of such a dependency tree is shown in Fig. 3, where the words represent the nodes of the tree and the arcs the dependencies between words. Dependencies are oriented, with each arrow pointing to the dependent word of the relation. Each dependency is further labelled with the name of a syntactic relation, such as *Sb* for subject, *Pred* for predicate, *Atr* for attribute, *Obj* for object, etc.

Dependency label The first generic feature derived from the parse tree is the label of the dependency from the target word to its head. For example, the values taken by this feature in Fig. 3 are, for each word: *Atr, Sb, AuxP, Atr, Root, Pnom, Obj.*

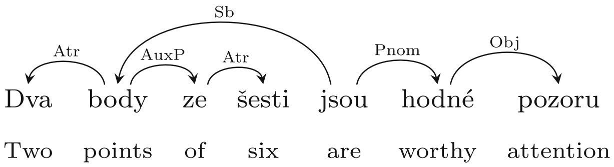

Sb
Atr
AuxP
Atr
Pnom
Obj
Dva body ze šesti jsou hodné pozoru
Two points of six are worthy attention

Fig. 3 Example of an instance of a Czech *statement* dialogue act (*two points out of six are worthy of attention*) with its parse tree

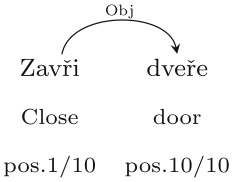


Fig. 4 Example of leftmost root position

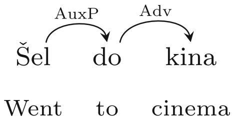


Fig. 5 Example of unexpressed subject

Root position in the utterance In theory, every utterance is parsed into a single dependency tree. The position of the root of this tree is likely to depend on the type of sentence and dialogue act. Hence, intuitively, the root tends to be positioned in the middle of declarative sentences, as in Fig. 3, while it is more often located at the start of utterances for commands/orders, such as in: “Zavři dveře!”(Close the door!) (Fig. 4).

This feature is the absolute position of the root, after normalization of the sentence length to 10 words. The normalization is realized with a standard binning technique, eventually filling empty internal bins with virtual non-root words for short sentences, so that the word at the middle of the sentence is in bin 5 and recursively in the left and right halves of the sentence.

Unexpressed subject This feature is a boolean feature that is true if and only if a subject dependency exists for the first verb in the sentence. Indeed, verbs without subjects may intuitively occur more frequently in commands/orders than in declarative sentences, as illustrated in the previous example. This is however not always true, especially in the Czech language, where unexpressed subjects are quite common and thus often occur in most dialogue acts, such as in: “Šel do kina.” (He went to the cinema.) (Fig. 5).

Basic composite pair This feature is a boolean value that encodes the relative position of each pair *Subject* and *verb*. When the verb precedes the subject, this is often viewed as strong evidence in favour of detecting a question in many European languages such as English and French. However, in the Czech language, this is not always true because of two main factors:

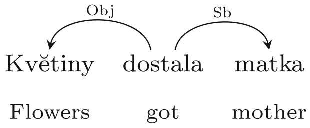


Fig. 6 Example of inverted subject

1. Subjects may be omitted, as explained in the previous section.
2. A statement can start with a *Direct Object*, followed by a *Verb* and its *Subject*, such as in "Květiny dostala matka." (The mother got flowers) (Fig. 6).

4 Dialogue act model

4.1 General principle

The general principle of our dialogue act recognition approach is to decompose the problem of tagging a complete sentence into the (easier) problems of tagging individual words. Our basic assumption is that every single word contributes to the global dialogue act depending on its form, nature and global context. The proposed approach thus assigns a single dialogue act tag to every word and then combines all the dialogue act tags that contribute to the same sentence to infer a single dialogue act tag for this sentence. The word-tagging process is implemented with a conditional random field (CRF) while the sentence-tagging process is realized with two simple combination models that are described next.

4.2 Training and pre-processing

Only the word-level CRF model is trained. The second combination stage is realized by a non-parametric decision process and thus does not need any training.

The manually annotated dialogue act tag associated to each training utterance is first duplicated and assigned to every word of the utterance. Then, these utterances are automatically tagged with POS-tags and parsed with the Malt parser (Nivre et al. 2007) to produce a dependency tree. A vector of lexical and syntactic features is then derived from this parse tree for each word of the utterance. A special word is inserted before every utterance, with a single feature that indicates the start of an utterance. This special word is given the same dialogue act tag as the other words of the sentence. Finally, all these feature vectors, along with their associated dialogue act tags are pooled together in sequence and the CRF is trained on this corpus with the classical L-BFGS algorithm.

The data pre-processing procedure described above also applies to the test corpus.

4.3 Testing and dialogue act inference

During testing, both word-level and sentence-level models are involved to infer dialogue acts. In the first step, the previously trained CRF is applied on the current words sequence and outputs one dialogue act tag for every word of the sentence. Then, the sentence-level decision process converts this resulting sequence of dialogue act tags into a single dialogue act tag per sentence.

Note that an alternative, single-stage strategy may have been to use a non-stochastic global approach, for instance with a maximum entropy model and global features. However, such an approach usually exploits a bag-of-word hypothesis or otherwise implies to explicitly define sentence-global features. Although we have already used with some success a similar approach in a previous work with words position (Král et al. 2007), we rather investigate in the current work the proposed two-stage strategy, which focuses on modelling the succession of word-level dialogue act tags.

Hidden Markov Models, Maximum-Entropy Markov Models (MEMMs) and CRFs are amongst the most common stochastic classifiers. We have chosen CRF because Laferty et al. (2001) have shown that CRFs avoid the label bias problem, as compared to MEMMs. Furthermore, CRFs are conditional models, and as such, make a better use of their parameters than generative models such as HMMs to model the target distribution of probability. They have also proven in recent years to be superior to most variants of HMMs in many natural language processing tasks and in particular in a punctuation generation application, which is closely related to dialogue act recognition. Hence, Favre et al. (2009) compared three sequence models: Hidden-Event Language Model (HELM), factored-HELM and CRFs for comma prediction. They have shown that the best results are obtained with CRFs, although CRFs may not scale easily to large databases.

We thus use CRFs to compute the conditional probability:

$$\begin{aligned} P(DA|F) &= P(c_0, c_1, \ldots, c_n | f_0, f_1, \ldots, f_n) \\ &= \prod_{i=0}^{n} P(c_i | f_i, c_1, \ldots, c_{i-1}) \end{aligned} \tag{1}$$

where $F = \langle f_0, f_1, \ldots, f_n \rangle$ represents the sequence of features vectors, n is the number of words in the utterance, f_0 the initial start word and $DA = \langle c_0, c_1, \ldots, c_n \rangle$ is the output sequence of dialogue acts.

4.4 Sentence-level combination and decision process

We investigate two approaches for the final decision process, which shall output a single dialogue act tag for the whole utterance: majority voting and Naive Bayes classification.

4.4.1 Majority voting

The final dialogue act tag is simply the tag with the highest frequency counts amongst the n tags $c_1, \ldots, c_n$. Ambiguous cases are resolved by choosing the tag with the largest posterior probability.

4.4.2 *Naive Bayes classification*

In the "Naive Bayes" classifier (Grau et al. 2004), every tag c_i is assumed independent from all others given the Markov assumption. Hence, the probability over the whole utterance is given by Eq. 2:

$$P(c|F) = \prod_{i=1}^{n} P(c_i = c|f_i, c_{i-1}) \tag{2}$$

where $P(c_i = c|f_i, c_{i-1})$ is the word-level posterior probability returned by the first-order CRF, when the CRF is constrained to follow the sub-optimal path $(c_0 = c, c_1 = c, \ldots, c_n = c)$. Note that this tag sequence is different from the one used in the "majority voting case", where the global optimal path returned by the CRF is used.

The resulting dialogue act is the one that maximizes the a posteriori probability:

$$\hat{c} = \arg\max_c P(c|F).$$

5 Evaluation

The proposed two-step model is evaluated on a Czech train reservation corpus and compared with a unigram model and with a baseline CRF model that only exploits lexical and morpho-syntactic features. The evaluation metric is the dialogue act recognition accuracy. In the following, we first describe the Czech corpus, then the two pieces of software that have been used to compute the morphosyntactic tags and the parse tree and we finally discuss the experimental results.

5.1 Corpus

The corpus used to validate the proposed approaches is the Czech Railways corpus that contains human–human dialogues. It was created at the University of West Bohemia mainly by members of the Department of Computer Science and Engineering in the context of a train ticket reservation dialogue expert system. The whole corpus has been recorded in laboratory conditions and contains about 12 h of audio recordings. The audio files have been both manually and automatically transcribed. We thus evaluate our dialogue act recognition approach on both types of transcriptions, in order to further assess its robustness to speech recognition errors.

Automatic transcription has been realized with the jLASER (Pavelka and Ekštein 2007) recogniser, which has been developed in our LICS[3] laboratory. It is based on a so called *hybrid* framework that combines the advantages of the hidden Markov model approach with those of artificial neural networks. We use HMMs with state emission probabilities computed from the output neuron activations of a neural network (such as the multi-layer perceptron). jLASER has been trained on 6,234

[3] http://liks.fav.zcu.cz.

sentences (about 9 h), while 2,173 sentences (about 3 h) pronounced by different speakers are used for testing. Because of the size of the corpus, a class-based 3-g language model has been used.

All sentences of this "test" corpus have been manually labelled by three different annotators with the following dialogue acts: statements (S), orders (O), yes/no questions [Q (y/n)] and other questions (Q). The DA corpus structure is reported in Table 2, where the number of dialogue acts is shown in column 2. This choice of dialogue acts has been done because our DA recognition module is designed to be used with a rule-based dialogue system that only exploits these four types of information as an input. The following dialogue act recognition experiments are realized on this labelled corpus using a cross-validation procedure, where 10 % of the corpus is reserved for the test, another 10 % for the development set and 80 % for training of the CRF.

5.2 Tools

For lemmatization and POS-tagging, we use the mate-tools http://code.google.com/p/mate-tools/. The lemmatizer and POS tagger models are trained on 5,853 sentences (94,141 words) randomly taken from the Prague Dependency Tree Bank (PDT 2.0) (Hajič et al. 2000) corpus. The PDT 2.0 is a collection of Czech newspaper texts that are annotated on the three following layers: morphological (2

Table 2 Description of the Czech Railways DA corpus

DA corpus			
DA	No.	Example	English translation
S	566	Chtěl bych jet do Písku	I would like to go to Písek
O	125	Najdi další vlak do Plzně!	Give me the next train to Plzeň!
Q [y/n]	282	Řekl byste nám další spojení?	Do you say next connection?
Q	1,200	Jak se dostanu do Šumperka?	How can I go to Šumperk?
Sent.	2,173		

Table 3 Tag-set description

Abbreviation	Description
A	Adjective
C	Numeral
D	Adverb
I	Interjection
J	Conjunction
N	Noun
P	Pronoun
V	Verb
R	Preposition
T	Particle
Z	Punctuation

million words), syntactic (1.5 million words) and complex syntactic and semantic layer (0.8 million words). In this work, only the syntactic dependencies of the second layer are considered. The performance of the lemmatizer and POS tagger are evaluated on a different set of 5,181 sentences (94,845 words) extracted from the same corpus. The accuracy of the lemmatizer is 81.09 %, while the accuracy of our POS tagger is 99.99 %. Our tag set contains 11 POS-tags as described in Table 3.

Our dependency parser is the Malt Parser v 1.3 trained on 32,616 sentences (567,384 words) from PDT 2.0. The dependency set is thus: *Adv, AdvAtr, Apos, Atr, AtrAdv, AtrAtr, AtrObj, Atv, AtvV, AuxC, AuxG, AuxK, AuxO, AuxP, AuxR, AuxT, AuxV, AuxX, AuxY, AuxZ, Coord, ExD, Obj, ObjAtr, Pnom, Pred, Sb*. The Labelled Attachment Score (LAS) of our parser is about 66 %.

Our CRF toolkit is based on the Stanford OpenNLP library,[4] which has been modified in order to include syntactic features. The resulting model has about 3,200 parameters.

5.3 Baseline rule-based system

Our claim in this work is that structured syntactic features, which cannot be simply derived from word forms, bring relevant information that help a classifier to discriminate between some dialogue acts, even in Czech, which is known to be a free-word order language. We actually show next that despite the theoretical linguistic constructions in Czech, which do not a priori strongly constrain the grammatical structures with regard to word orders, common usage in Czech exhibits statistical properties that are discriminative for the few dialogue acts considered here. Furthermore, we show that such statistical properties cannot be captured with simple deterministic rules, but that they must be considered instead in context within a stochastic model like the proposed CRF that is trained on real data.

To illustrate this idea, let's consider the particularly difficult case of yes–no questions versus statement. Table 4 shows a typical example of such a case, which cannot be captured by syntactic information in Czech.

However, despite such difficult theoretical constructions, we have automatically parsed our Czech speech corpus and analyzed the relative frequency of subject relations with the verb on the left of the subject (feature "Basic composite pair"): 48 % of such inverted relations occur in statements, which corresponds to a pure random ratio and complies with the free-word order property of Czech, while this ratio goes up to 88 % in yes/no questions, which demonstrates that such a feature is indeed informative in common usages of Czech. Nevertheless, we also show next that this observation, in itself, is not enough to accurately discriminate between yes/no questions and statements, and that it must be considered in context to be really useful.

In order to validate this claim, we build next a deterministic baseline model that classifies the four proposed dialogue acts using hand-crafted rules that:

[4] http://incubator.apache.org/opennlp.

Table 4 Example of discrimination between yes/no question and statement that cannot be realized from syntactic information in Czech

	English	Czech
Statement	He loves her	Miluješ ho
Yes/no question	Does he love her	Miluješ ho

Punctuation is not shown because it is not available at the output of speech recognizers and is thus not used by our system

Table 5 Hand-crafted rules used in the deterministic baseline system

Rule	Trigger DA	Description
1. Lexical rules		
R1	S	Occurrence of a word in the list
		“bych”: conditional form, first person singular
		“bychom”: conditional form, first person plural
		“jsem”: “to be”, first person singular
		“jsme”: *“to be”*, first person plural
		“potřebuji”, “potřebujeme”
		“I need”, *“we need”*
		“chci”, “chceme”, “chtěl”, “chtěla”, “chtěli”
		“I want”, “we want”, “he wanted”, “she wanted”, “they wanted”
		“znát”, “vědět” and “doptat”
		“to know” and “to ask”
R2	Q [y/n]	The first word in the sentence is one of
		“můžu”, “můžete”, “můžeme”
		“can you” in singular and in plural, *“can we”*
		“má”, “máte”, “máme”
		“do you have” in singular and in plural, *“do we have”*
R3	Q	“Wh*” word or a word from the list below is at the beginning of the sentence
		“jak”, “jakpak” and “kolik”
		“how” and *“how many”*
2. Morpho-syntactic rules		
R4	O	Verb with imperative form at the beginning of the sentence
		Verb at the beginning of the sentence has a suffix amongst
		“ej”, “me”, “te”
		Suffix of imperative form for 2nd person singular, 1st person plural, 2nd person plural
3. Syntactic rules		
R6	O	The first word of the sentence is the syntactic root
R7	O	A verb doesn’t have any subject
R8	Q [y/n]	Subject–verb inversion
4. Default rule		
R9	S	When no previous rules apply, by default, the statement is chosen

Table 6 Dialogue act recognition accuracy for the baseline rule-based system with manual and automatic word transcription by jLASER recognizer

Transcription type	S	O	Q [y/n]	Q	Global
Manual	67.3	95.2	92.2	87.9	83.5
jLASER	67.5	88.8	85.1	82.0	79.0

- Include common lexical knowledge, such as interrogative words.
- Use syntactic rules that match the proposed features described in Sect. 3.1.2, such as the subject–verb inversion rule just described.

The set of rules is described in Table 5. When several rules apply on the same sentence, the chosen dialogue act is decided with a majority vote. In case of equality, the winner amongst competing dialogue acts is the one with the higher prior probability on the corpus, i.e., in decreasing order: Q, S, Q [y/n], O.

The recognition accuracy of the rule-based system is shown in Table 6. We evaluate two cases: manual word transcription and automatic transcription by jLASER recognizer. Table 6 shows that errors from the speech recognizer don't play an important role for DA recognition, resulting in a decrease of accuracy of about 4 %.

The highest score for class O might result from the precision of the set of rules defined for this class. Conversely, the lower score of class S may be due to the difficulty to define a specific rule for this class. We thus only used a list of keywords for S and sentences are mainly classified into this class when no rule from another class is triggered.

5.4 Experiments

Two experiments are realized next. The first one performs dialogue act recognition on manual word transcriptions and evaluates and compares the impact of the proposed lexical and syntactic features and the relative performances of both sentence-level combination models. The unigram model corresponds to a very basic baseline approach that only exploits lexical unigram probabilities. We further compare the proposed approach with more advanced baselines that are also based on a CRF but with lexical and morphosyntactic features only (*word forms*, *lemmas* and *POS-tags*). In the second experiment, the same set of models is applied on automatic word transcriptions. This allows assessing the robustness of both our parsers and feature sets to speech recognition errors.

5.4.1 Manual transcription evaluation

Table 7 shows the dialogue act recognition accuracies obtained with the different proposed models. We have computed statistical significance of the difference between two models with the McNemar test, as suggested in Gillick and Cox (1989) for a similar classification task. The p value is in general below the traditional threshold of 0.05. The p values of some important comparisons are for instance:

Table 7 Dialogue act recognition accuracy for different features/approaches with manual word transcription

	Features/approach	Accuracy (%)				
		S	O	Q [y/n]	Q	Global
1. Unigram						
B0	Words	93.5	77.6	96.5	89.9	91.0
2. Majority voting						
B1MV	Words	87.63	76.61	81.21	99.42	92.50
B2MV	Words + lemmas	87.63	76.61	82.27	99.42	92.82
B3MV	Words + POS-tags	87.63	72.58	81.21	99.50	92.50
B4MV	Words + pronoun at the beginning	90.28	76.61	95.39	99.33	95.17
B5MV	Words + verb at the beginning	87.63	79.03	95.04	99.42	94.61
BAMV	Words + all baseline features	88.87	84.68	94.68	99.42	95.21
S1MV	Words + dependency labels	87.81	74.19	88.30	99.42	93.51
S2MV	Words + root position	89.05	87.10	92.20	99.33	95.03
S3MV	Words + unexpressed subject	89.22	78.23	84.75	99.25	93.55
S4MV	Words + basic composite pair	88.34	78.23	84.75	99.33	93.37
SyMV	All features	89.93	92.74	94.68	99.42	95.95
3. Naive Bayes classifier						
B1NB	Words	94.52	71.77	83.33	99.25	94.38
B2NB	Words + lemmas	94.88	87.90	91.13	99.42	96.50
B3NB	Words + POS-tags	95.05	80.65	88.30	99.00	95.53
B4NB	Words + pronoun at the beginning	96.47	83.06	94.68	99.33	97.05
B5NB	Words + verb at the beginning	88.87	48.39	94.33	99.00	92.86
BANB	Words + all baseline features	92.05	86.29	94.68	99.5	96.18
S1NB	Words + dependency labels	94.52	81.45	89.01	99.00	95.53
S2NB	Words + root position	90.11	85.48	93.62	99.00	95.21
S3NB	Words + unexpressed subject	93.11	85.48	85.82	99.08	95.03
S4NB	Words + basic composite pair	93.11	85.48	86.52	99.08	95.12
SyNB	All features	95.23	95.16	96.81	99.33	97.70

- SyNB versus B3NB: $p < 0.001$.
- SyNB versus B4NB: $p = 0.016$.
- SyNB versus BANB: $p = 0.002$.

We can first observe that the Naive Bayes combination gives in general better results than majority voting, which was expected, as Naive Bayes exploits the posteriors, which are a richer source of information than just the knowledge of the winning class.

This table also shows relatively low recognition scores for the class O. This is probably due to the relatively smaller amount of training data for this class. This analysis is supported by the good recognition accuracy obtained by the baseline rule-based system for this class, which does not depend on any training corpus. The best recognition rate is for the class Q, which is both the most frequent class and which is characterized by strong cues, especially concerning the influence of the first word in the sentence (B4NB) as well as distinctive interrogative word forms (B1NB, B2NB).

The most important remark is that the combination of all proposed syntactic and baseline features significantly outperforms all baseline features, which confirms that the proposed syntactic features bring complementary information. This result supports our claim that structured syntactic information might prove useful for dialogue act recognition.

5.4.2 Automatic transcription evaluation

Table 8 shows a similar evaluation to the one in Table 7, except that the textual transcriptions are now obtained automatically with the jLASER speech recogniser.

Table 8 Dialogue act recognition accuracy for different features/approaches with automatic word transcription using jLASER speech recogniser

	Features/approach	Accuracy (%)				
		S	O	Q (y/n)	Q	Global
1. Unigram						
B0	Words	93.1	68.8	94.7	86.3	88.2
2. Majority voting						
B1MV	Words	82.69	29.84	62.06	99.25	86.14
B2MV	Words + lemmas	83.92	41.13	65.25	99.17	87.48
B3MV	Words + POS-tags	81.80	28.23	63.48	99.17	85.96
B4MV	Words + pronoun at the beginning	86.04	54.03	80.50	98.75	90.52
B5MV	Words + verb at the beginning	81.10	25.00	79.43	98.17	87.11
BAMV	Words + all baseline features	88.52	58.47	81.25	97.97	90.84
S1MV	Words + dependency labels	80.04	33.06	70.57	99.25	86.74
S2MV	Words + root position	83.39	50.00	79.08	98.33	89.18
S3MV	Words + unexpressed subject	81.63	33.87	61.35	99.33	86.05
S4MV	Words + basic composite pair	81.63	33.87	60.28	99.33	85.91
SyMV	All features	84.28	71.77	82.98	98.17	91.07
3. Naive Bayes classifier						
B1NB	Words	88.16	29.84	79.79	99.17	89.83
B2NB	Words + lemmas	90.81	33.87	82.27	99.33	91.16
B3NB	Words + POS-tags	89.93	30.65	80.85	99.08	90.42
B4NB	Words + pronoun at the beginning	91.34	66.13	89.36	98.67	93.21
B5NB	Words + verb at the beginning	82.51	29.03	82.62	97.83	87.94
BANB	Words + all baseline features	88.34	68.55	87.59	97.67	92.27
S1NB	Words + dependency labels	87.99	33.06	81.56	99.25	90.24
S2NB	Words + root position	84.81	69.35	83.33	88.42	91.25
S3NB	Words + unexpressed subject	89.93	34.68	80.50	99.08	90.61
S4NB	Words + basic composite pair	89.75	33.06	80.14	99.17	90.47
SyNB	All features	91.17	70.97	88.65	98.00	93.46

Sentence recognition accuracy is 39.8 % and word recognition accuracy is 83.4 %. The complete annotation process starts from these imperfect transcriptions, including: lemmatization, POS-tagging, parsing and dialogue act recognition. This experiment thus assess the robustness of the complete processing chain to speech recognition errors, in order to match as closely as possible the actual use of the proposed approach in realistic conditions.

We can first observe that the impact of speech recognition errors is moderately large, but not dramatic and thus does not jeopardize the applicability of the proposed approach in real conditions. Hence, while the dialogue act classification errors increase by 30 % with the unigram model, they increase by 113 % with the baseline CRF B3NB, which was expected because the CRF exploits the correlation between successive words and tags, which may propagate errors amongst words. However, despite its lower robustness, the CRF model still performs better in absolute value than the unigram model. The increase in classification error of the syntactic-aware model is about 183 %, which is due to the greater sensibility of the processing chain for this model. Indeed, speech recognition errors are known to have a large impact on POS-tagging and parsing performances. The derived syntactic features are thus also largely impacted by such errors. This also explains why the simple proposed baseline features, such as B4NB, are also the most robust ones.

6 Conclusions

This work extends our previous works that tended to demonstrate the importance of global structural information for dialogue act recognition by implicitly modelling local constraints with CRFs and explicitly proposing global syntactic features derived from automatic parsing of the sentence. Regarding the efficiency of syntactic features for dialogue act recognition, we have provided a number of evidence to support our claim that syntactic information might be important for dialogue act recognition and that the main reason why they have not been widely used so far in this domain is due to (1) the difficulty to reliably parse speech and dialogues; (2) the intrinsic complexity of the syntactic material as compared to the classical lexical and morphosyntactic tags; and (3) the lack of robustness of parsers to speech recognition errors. This claim is based on a review of several companion works that show the importance of syntax for both dialogue act recognition and closely related domains such as punctuation generation. Second, we have proposed several simple as well as more complex syntactic features that are derived from a full deep parsing of the sentence and have shown that the use of such features indeed significantly improves the dialogue act classification performance on our Czech corpus. Finally, we have studied the robustness of the proposed system and have shown that, as expected, the most complex syntactic features are also the most sensitive to speech recognition errors.

Hence, given the evidence collected in this work, we conclude that syntax information might prove important for dialogue act recognition, as it has already been shown relevant for many other natural language processing tasks. The main challenge that remains is to increase its robustness to speech recognition errors, but we expect this challenge to be soon overcome, thanks to the great progresses realized in the automatic parsing community in recent years.

Acknowledgments This work has been partly supported by the European Regional Development Fund (ERDF), project "NTIS—New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090. We would like also to thank Ms. Michala Beranová for some implementation work.

References

Alexandersson, J., Reithinger, N., & Maier, E. (1997). *Insights into the dialogue processing of VERBMOBIL*. Tech. rep. 191, Germany: Saarbrücken.

Allen, J., & Core, M. (1997). Draft of DAMSL: Dialog act markup in several layers. http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html.

Andernach, T. (1996) *A machine learning approach to the classification of dialogue utterances*. Computing Research Repository.

Ang, J., Liu, Y., & Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the ICASSP*, Philadelphia, USA.

Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon Press.

Bilmes, J. (2005). *Backoff model training using partially observed data: Application to dialog act tagging*. Tech. rep. UWEETR-2005-0008, Department of Electrical Engineering, University of Washington.

Blanchon, H., & Boitet, C. (2000). Speech translation for French within the C-STAR II consortium and future perspectives. In *INTERSPEECH '00* (pp. 412–417).

Bunt, H. (1994). Context and dialogue control. *Think Quarterly, 3*, 19–31.

Carberry, S. (1990). *Plan recognition in natural language dialogue*. Cambridge, MA: MIT Press.

Cerisara, C., Král, P., & Gardent, C. (2011). Commas recovery with syntactic features in French and in Czech. In *INTERSPEECH'11* (pp. 1413–1416), Firenze, Italy.

Crook, N., Granell, R., & Pulman, S. (2009). Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Proceedings of the 10th annual meeting of the special interest group in discourse and dialogue (SIGDIAL)* (pp. 241–348).

Dhillon, R. B. S., & Carvey, H. S. E. (2004). *Meeting recorder project: Dialog act labeling guide*. Tech. rep. TR-04-002, International Computer Science Institute.

Di Eugenio, B., Xie, Z., & Serafin, R. (2010). Dialogue act classification, higher order dialogue structure, and instance-based learning. *Journal of Discourse and Dialogue Research, 1*(2), 1–24.

Dielmann, A., & Renals, S. (2008). Recognition of dialogue acts in multiparty meetings using a switching DBN. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(7), 1303–1314.

Favre, B., Hakkani-Tür, D., & Shriberg, E. (2009). Syntactically-informed models for comma prediction. In *ICASSP '09* (pp. 4697–4700), Taipei, Taiwan.

Garner, P. N., Browning, S. R., Moore, R. K., & Russel, R. J. (1996). A theory of word frequencies and its application to dialogue move recognition. In *ICSLP '96* (Vol. 3, pp. 1880–1883), Philadelphia, USA.

Geertzen, J. (2009). *Dialog act recognition and prediction*. Ph.D. thesis, University of Tilburg.

Gillick, L., Cox, S. (1989). Some statistical issues in the comparison of speech recognition algorithms. In *ICASSP '1989* (pp. 532–535).

Grau, S., Sanchis, E., Castro, M. J., & Vilar, D. (2004). Dialogue act classification using a Bayesian approach. In *9th international conference speech and computer (SPECOM '2004)* (pp. 495–499), Saint-Petersburg, Russia.

Guo, Y., Wang, H., & Genabith, J. V. (2010). A linguistically inspired statistical model for Chinese punctuation generation. *ACM Transactions on Asian Language Information Processing, 9*(2), 27.

Hajičová, E. (2000). *Dependency-based underlying-structure tagging of a very large Czech corpus*, *41*(1), 57–78.

Hajič, J., Böhmová, A., Hajičová, E., & Vidová-Hladká, B. (2000). The Prague dependency treebank: A three-level annotation scenario. In A. Abeillé (Ed.), *Treebanks: Building and using parsed corpora* (pp. 103–127). Amsterdam: Kluwer.

Jekat, S., et al. (1995). *Dialogue acts in VERBMOBIL*. Verbmobil report 65.

Jeong, M., & Lee, G. G. (2008). Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(7), 1287–1302.

Ji, G., & Bilmes, J. (2005). Dialog act tagging using graphical models. In *Proceedings of the ICASSP* (Vol. 1, pp. 33–36), Philadelphia, USA.

Joty, S., Carenini, G., & Lin, C.-Y. (2011). Unsupervised approaches for dialog act modeling of asynchronous conversations. In *Proceedings of the IJCAI*, Barcelona, Spain.

Jurafsky, D., et al. (1997). Automatic detection of discourse structure for speech recognition and understanding. In *IEEE workshop on speech recognition and understanding*, Santa Barbara.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics* (2nd ed.). Upper Saddle River: Prentice-Hall.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBD–DAMSL shallow-discourse-function annotation* (Coders manual, draft 13). Tech. rep. 97-01, University of Colorado, Institute of Cognitive Science.

Kautz, H. A. (1987). *A formal theory of plan recognition*. Tech. rep. 215. NY: Department of Computer Science, University of Rochester.

Keizer, S. A. R., & Nijholt, A. (2002). Dialogue act recognition with Bayesian networks for Dutch dialogues. In *3rd ACL/SIGdial workshop on discourse and dialogue* (pp. 88–94), Philadelphia, USA.

Klüwer, T., Uszkoreit, H., & Xu, F. (2010). Using syntactic and semantic based relations for dialogue act recognition. In *Proceedings of the 23rd international conference on computational linguistics: Posters (COLING '10)* (pp. 570–578). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://portal.acm.org/citation.cfm?id=1944566.1944631.

Kompe, R. (1997). *Prosody in speech understanding systems*. Berlin: Springer.

Král, P., Cerisara, C., & Klečková, J. (2005). Combination of classifiers for automatic recognition of dialog acts. In *Interspeech '2005* (pp. 825–828). Lisboa, Portugal: ISCA.

Král, P., Cerisara, C., & Klečková, J. (2006a). *Automatic dialog acts recognition based on sentence structure*. In *ICASSP '06* (pp. 61–64), Toulouse, France.

Král, P., Klečková, J., Pavelka, T., & Cerisara, C. (2006b). Sentence structure for dialog act recognition in Czech. In *ICTTA '06*, Damascus, Syria.

Král, P., Cerisara, C., & Klečková, J. (2007). Lexical structure for dialogue act recognition. *Journal of Multimedia (JMM), 2*(3), 1–8.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the eighteenth international conference on machine learning (ICML '01) (pp. 282–289). San Francisco, CA: Morgan Kaufmann. URL: http://portal.acm.org/citation.cfm?id=645530.655813.

Lavie, A., Pianesi, F., & Levin, L. (2006). The NESPOLE! System for multilingual speech communication over the internet. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(5), 1664–1673.

Lendvai, P. A., & van den Bosch, K. E. (2003). Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In *Workshop on dialogue systems: Interaction, adaptation and styles management (EACL-03)* (pp. 69–78). Hungary: Budapest.

Levin, L., Langley, C., Lavie, A., Gates, D., Wallace, D., & Peterson, K. (2003). Domain specific speech acts for spoken language translation. In *4th SIGdial workshop on discourse and dialogue*. Japan: Sapporo.

Litman, D. J. (1985). *Plan recognition and discourse analysis: An integrated approach for understanding dialogues*. Ph.D. thesis, Rochester, NY: University. of Rochester.

Mast, M., et al. (1996). Automatic classification of dialog acts with semantic classification trees and polygrams. In *Connectionist, statistical and symbolic approaches to learning for natural language processing* (pp. 217–229).

Mast, M., Kompe, R., Harbeck, S., Kiessling, A., Niemann, H., Nöth, E., et al. (1996). Dialog act classification with the help of prosody. In *ICSLP '96*, Philadelphia, USA.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., et al. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering, 13*(2), 95–135.

Orkin, J., & Roy, D. (2010). Semi-automated dialogue act classification for situated social agents in games. In *Proceedings of the agents for games and simulations workshop at the 9th international conference on autonomous agents and multiagent systems (AAMAS)*, Toronto, Canada.

Pavelka, T., Ekštein, K. (2007). JLASER: An automatic speech recognizer written in Java. In *XII international conference speech and computer (SPECOM '2007)* (pp. 165–169), Moscow, Russia.

Petukhova, V., & Bunt, H. (2011). Incremental dialogue act understanding. In *Proceedings of the 9th international conference on computational semantics (IWCS-9)*, Oxford.

Power, R. J. D. (1979). The organization of purposeful dialogues. *Linguistics, 17*, 107–152.

Quarteroni, S., Ivanov, A. V., & Riccardi, G. (2011). Simultaneous dialog act segmentation and classification from human–human spoken conversations. In *Proceedings of the ICASSP*, Prague, Czech Republic.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest semantics for the organization of turn-taking in conversation. *Language, 50*(4), 696–735.

Samuel, K., Carberry, S., & Vijay-Shanker, K. (1998). Dialogue act tagging with transformation-based learning. In *17th international conference on computational linguistics* (Vol. 2, pp. 1150–1156). Morristown, NJ, USA, Montreal, QC, Canada: Association for Computational Linguistics.

Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist, 70*(1), 1075–1095.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*.

Serafin, R., & Di Eugenio, B. (2004). LSA: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Spain*.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., et al. (1998). Language and speech, Vol. 41 of special double issue on prosody and conversation, Ch. can prosody aid the automatic classification of dialog acts in conversational speech? (pp. 439–487).

Sporleder, C., & Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: A critical assessment, *Natural Language Engineering, 14*(3).

Stolcke, A. et al. (2000). Dialog act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics, 26*, 339–373.

Traum, D. R. (1999). Speech acts for dialogue agents. In M. Wooldridge & A. Rao (Eds.), *Foundations and theories of rational agents*. (pp. 169–201). Dordrecht: Kluwer.

Tur, G., Guz, U., & Hakkani-Tur, D. (2006). Model adaptation for dialogue act tagging. In *Proceedings of the IEEE spoken language technology workshop*.

Verbree, D., Rienks, R., & Heylen, D. (2006). Dialog-act tagging using smart feature selection; results on multiple corpora. In *The first international IEEE workshop on spoken language technology (SLT)*, Aruba, Palm Beach.

Webb, N. (2010). *Cue-based dialog act classification*, Ph.D. thesis, University of Sheffield.

Wright, H. (1998). Automatic utterance type detection using suprasegmental features. In *ICSLP '98* (Vol. 4), Sydney, Australia.

Wright, H., Poesio, M., & Isard, S. (1999). Using high level dialogue information for dialogue act recognition using prosodic features. In *ESCA workshop on prosody and dialogue*, Holland, Eindhoven.

Zhou, K., & Zong, C. (2009). Dialog-act recognition using discourse and sentence structure information. In *Proceedings of the 2009 international conference on asian language processing (IALP '09)* (pp. 11–16). Washington, DC, USA: IEEE Computer Society.

Zimmermann, M., Stolcke, A., & Shriberg, E. (2006). Joint segmentation and classification of dialog acts in multiparty meetings. In *ICASSP '06* (pp. 581–584), Toulouse, France.

Commas recovery with syntactic features in French and in Czech

Christophe Cerisara* Pavel Král+ Claire Gardent*

* LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
France

+Dept. of Computer Science & Engineering
University of West Bohemia
Plzeň, Czech Republic

Abstract

Automatic speech transcripts can be made more readable and useful for further processing by enriching them with punctuation marks and other meta-linguisitic information. We study in this work how to improve automatic recovery of one of the most difficult punctuation marks, commas, in French and in Czech. We show that commas detection performances are largely improved in both languages by integrating into our baseline Conditional Random Field model syntactic features derived from dependency structures. We further study the relative impact of language-independent vs. specific features, and show that a combination of both of them gives the largest improvement. Robustness of these features to speech recognition errors is finally discussed.

Index Terms: Dependency parsing, punctuation detection, commas recovery

1. Introduction and related works

Automatic speech transcripts are still difficult to read, because of recognition errors, but also because of the missing structure of the document, and in particular capitalization and punctuation. We focus in this work on the task of recovering commas in a given text, which may also help subsequent automatic processing such as parsing and mining.

Punctuation recovery is often realized based on prosodic (pauses, pitch contours, energy) and lexical (surrounding words, n-grams) features, such as in [1], where full stops, commas and question marks are recovered using a finite state approach that combines lexical n-grams and prosodic features. Commas are recovered with a Slot Error Rate (SER) of 81% on automatically transcribed utterances of the Hub-4 English audio corpus. Both prosodic and lexical features are also combined via a maximum entropy model in [2], where commas are recovered on the Switchboard corpus with a F-score of 79% with lexical features only, while prosody does not help at all. For English, both the works reported in [3] and [4] (described next) show that syntactic features are very important for punctuation recovery.

Punctuation recovery has also been studied in other languages than English: In [5], automatic capitalization is realized along with automatic recovery of full stops and commas in Portuguese. Both punctuation marks are detected with a maximum entropy model that exploits acoustic and lexical features. Commas are hence recovered on automatic speech transcripts with an SER of 101%.

The authors of [6] exploit a hidden-event n-gram model combined with a prosodic model to recover punctuation marks on the Czech broadcast news corpus. F-scores of 66% and 68% are reported for commas recovery with respectively the lexical n-gram only and the lexical model combined with the decision tree model for prosody. In both previous works, no syntactic features are used.

In [7], a maximum entropy model is also exploited to recover 14 punctuation marks from the Penn Chinese TreeBank. For commas, the model exhibits a F-score of 81.14%. In order to achieve such performances, syntactic features derived from the manual syntactic annotations are used.

The authors of [4] focus on the study of comma prediction in English with syntactic features. They have compared three sequence models: Hidden-Event Language Model (HELM), factored-HELM and Conditional Random Fields (CRF). They report that the best results have been obtained with CRF, although CRFs may not scale easily to large databases.

2. Commas recovery approach

We propose to extend the work of [4] in the following aspects:

- Design of new syntactic features dedicated to comma recovery and derived from dependency structures.
- Evaluation of these features on two new languages: French and Czech.

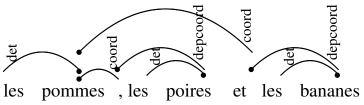
In French and Czech, the available corpora are far from being as large as in English, and scaling is not yet an issue. We have thus decided to base our work on CRF models. Furthermore, considering the relatively limited impact of prosodic features for commas recovery as reported in the literature, only lexical and syntactic features are exploited next. A CRF model is then trained to classify every subsequent word into two classes: the class of words that are followed by a comma, and the class of words without comma. The CRF input features are only local and derived from the current, previous and next words. These features are then pushed in sequence, with special words inserted at sentence boundaries, into a feature stream that is used to train the CRF model. The test corpus is processed in the same way.

3. Syntactic features

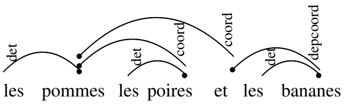
3.1. French dependency features

3.1.1. Syntactic feature for coordination

In French, commas are commonly used to serve different purposes [8]. One of their most common usage is as a replacement of coordinations, such as “et” (and) and “ou” (or). The following illustrates this usage of commas, for the nominal group “The apples, pears and bananas”:



The dependency tree is represented on top of the words, with oriented dependency arcs between the head word (circled extremity of the arc) and its governed word. Dependencies are labelled with grammatical functions, such as *det* for determiner, *coord* for coordinator and *depcoord* for coordination dependent. This example follows the annotation guidelines of the French Treebank, in which commas have an explicit role in the coordination structure. Our objective is to recover commas, which implies to remove them from the corpus first. This is achieved by automatically transforming the previous example tree into the following one, in order to preserve the coordination structure:

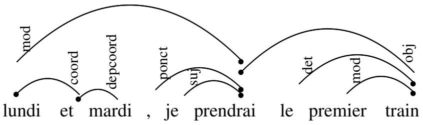


We have designed a feature to capture this usage of commas as follows: let w_i be the i^{th} word in the sentence. We want to know whether a comma shall be put between w_i and w_{i+1}. We then successively look at words w_{i+1}, w_{i+2}, $\cdots$ until we find a word w_{i+n} that is governed by a head word w_{i-m} $m \geq 0$ on the left of w_i. The feature is then TRUE iff this dependency label is "coord", otherwise, it is FALSE. This raw feature is further modified according to the identity of the following word w_{i+1}. When w_{i+1} is one of the two coordination keywords "et" (and) and "ou" (or), then the feature is set to FALSE. In the previous example, this feature is, for each word: FALSE (les), TRUE (pommes), TRUE (les), FALSE (poires), FALSE (et), FALSE (les), FALSE (bananes).

This feature is hereafter called **IsCoord**.

3.1.2. Syntactic feature for modifiers

Another common usage is to separate the modifier group that is before the verbal group, such as in "on Monday and Tuesday, I will take the first train":



We can note a few new dependencies in this example: *mod* for modifier, *ponct* for punctuation marks that are not part of a coordination structure, *suj* for subject and *obj* for object. In this example, *mardi* (Tuesday) should be followed by a comma, while *premier* (first) should not.

Intuitively, we will look at subtrees, or constituents, which are modifiers of another following word in the sentence. Then, commas may occur right after such constituents. This may be inferred, for the target word w_i, by looking for every subtree for which w_i is the rightmost word. Then, we check whether this subtree is a modifier of another word w_{i+m} $m > 0$ that is anywhere in the sentence after w_i. It is important to check that w_i is indeed the rightmost word of the modifier constituent, because commas usually only occur right after the constituent, and not within the constituent. Every time these conditions are met, the feature is defined as the "distance", i.e. the number of words between the head of this constituent and w_i.

In the previous example, "lundi" is the rightmost word of a single subtree composed of a single node (itself). The head of this subtree is thus also "lundi", which is indeed a modifier of a word at its right ("prendrai"). So the value of this feature for lundi is 0. Similarly, the feature value is 1 for "et" and 2 for "mardi". It is then -1 for "je", "prendrai" and "le", because there is no modifier, and it is 0 for "premier" and -1 for "train".

We hereafter refer to this feature as **IsMod**.

3.1.3. Syntactic feature for cross-dependencies

We propose here to generalize both previous features into a new feature that encodes cross-dependency relations between both parts of the sentence, before and after the target candidate word w_i. The intuitive idea behind this feature is that commas are more likely to separate two weakly dependent chunks than to occur within a chunk. This feature is computed as follows, for the target word w_i:

- We check whether the head of w_{i+1} is located before w_i; if so, then the corresponding dependency is crossing the limit between w_i and w_{i+1}. The value of the feature is then the label of this dependency.
- Otherwise, the same test is performed recursively for every ancestor of w_{i+1}, i.e., for the head of the head of w_{i+1}, and so on, until a crossing dependency is found or until the root of the tree is reached.
- If the root of the tree is reached without finding any crossing dependency, then we look recursively for a left-to-right crossing dependency on top of w_i, its head, etc.

In the previous example, the feature values are:
lundi(coord), et(depcoord), mardi(LEFTmod), je(LEFTsuj), prendrai(obj), le(obj), premier(obj), train(NIL)

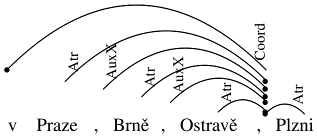
This feature is hereafter called **DepCross**. It is used next both in French and Czech experiments.

3.2. Czech dependency features

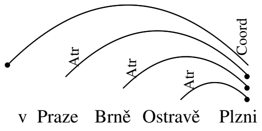
3.2.1. Syntactic feature for coordination

In the Czech language, the usage of commas is much more regular than in French, as described in [6], and commas most frequently precede specific grammatical function words, such as "protože" (because), "ale" (but), "který" (what), etc. This regularity mainly explains the relatively high F-scores reported in [6] and in our own experiments in Czech when using only contextual lexical features.

Nevertheless, another usage of commas in Czech that might be better recovered with syntactic features concerns, just like in French, coordination constructs. In the Prague Treebank, commas also play an explicit role in the syntactic tree, such as in "In Prague, Brno, Ostrava, Pilsen":



As we did for French, such structures are thus automatically transformed into:



A new syntactic feature has been specifically designed for this usage of commas in Czech. Intuitively, this feature aims at detecting coordination occurrences that involve more than three items. In such cases, just like in French, all items are usually separated by commas, except for the last two items that are separated by a coordination word, such as “a” (and) and “nebo” (or). The feature designed to detect this frequent pattern is computed as follow, for the target word w_i:

1. Recursively parse the tree branch from w_i to the root of the tree, i.e., the head of w_i, the head of its head, etc. until the root or a dependency ”coord” is found (*Plzni* in the example).
2. When such a coordinator is found, list all of its direct children and finds the dependency label that occurs the most frequently amongst them (*Atr*). We assume this is the dependency type between the coordinated items and their common head.
3. The children with this dependency type are sorted chronologically and all items that occur *after* the coordination keyword are removed (none in the example).
4. For each remaining item, we check whether w_i is the right-most word of the corresponding subtree (always true in the example, as every child is composed of a single-word); if it is, then the chosen feature is the number of remaining items between this target group and the coordination keyword (e.g., 2 for *Praze*).
5. The feature is set to -1 whenever any of the preceding conditions do not hold.

This feature is hereafter called **IsCoordCz**.

3.2.2. Dependency type feature

Apart from the special case of coordination, we have further used another generic feature derived from the parse tree, which is the label of the dependency from the target word to its head. This generic feature shall cover use cases for commas that are neither handled by lexical information only, nor by a coordination structure. For example, in the previous utterance “v Praze Brně Ostravě Plzni”, the features for each word are: *Root*, *Atr*, *Atr*, *Atr* and *Coord*.

This feature is hereafter called **DepLabel**. Just like **DepCross**, **DepLabel** has been tested on both French and Czech corpus, but it had no impact on the French F-score.

4. Experimental validation

4.1. Experimental set-up

The French Treebank (FTB) is composed of 12500 sentences and 325 000 words [9]. It consists of articles from *Le Monde* newspaper manually enriched with phrase structure annotations, which are further automatically converted into syntactic dependencies. The Prague Dependency Tree Bank (PDT 2.0) [10] is a collection of Czech newspaper texts that are annotated on the three following layers: morphological (2 million words), syntactic (1.5 million words) and complex syntactic and semantic layer (0.8 million words). In this work, only the syntactic dependencies of the second layer are considered. All experiments are realized in 10 folds cross-validation, where 9 tenths of the corpus is used to train the CRF model, and 1 tenth for testing.

We use a modified version of the Stanford CRF package initially developed for Named Entity Recognition [11] to train and test CRFs. The main modification concerns the possibility to complement the lexical and morphosyntactic features with the syntactic features for training and testing the CRF. The basic lexical features include (w_{i-1}, c), (w_i, c) and (w_{i+1}, c) where $c \in \{\text{comma}, \text{nocomma}\}$ is the class of word w_i. The morphosyntactic features (POS-tags) include (p_{i-1}, c), (p_i, c) and (p_{i+1}, c).

Two evaluation metrics are used: the classical F-score and the Slot Error Rate, as defined in [12].

4.2. Experimental results in French

Table 1 compares the performances of different feature sets for recovering commas on the French Treebank corpus, respectively with manual and automatic parses. The parsing is realized with our French version of the Malt Parser [13]. This parser has been trained on the very same 9 tenth of the corpus also reserved for training the comma-CRF. The CRF is actually always trained on the manual (gold) dependency trees of this 9 tenth corpus. Hence, automatic parsing is only used on the test set. Although this is a convenient approach because it does not require a double cross-validation procedure, it may be suboptimal, because the comma CRF only uses perfect dependency trees during training. On the other hand, the impact of parsing errors in table 1 is so small than it does not justify to train the CRF with parser errors.

	Manual deps.		Auto deps.	
	F-sc	SER	F-sc	SER
Lexical	38.7	93.6	38.7	93.6
Lexical + POS	43.2	85.4	**43.2**	85.4
Lexical + POS + IsCoord	46.9	82.7	46.5	83.0
Lexical + POS + IsMod	48.4	79.5	47.7	80.2
Lexical + POS + DepCross	75.0	50.7	74.9	51.1
Lexical + POS + all Synt.	76.4	47.6	**76.1**	48.4

Table 1: Comparison of different feature sets on the French Treebank. The confidence interval is $\pm 1.35\%$.

First, we can note that our baseline results are comparable to the baseline results of the state-of-the-art: hence our baseline F-score of 43.2% (for French) is comparable to the F-score of 46.9% (for English) obtained on the Gigaword corpus with lexical features in [4]. Note that there is a very large difference in corpus size between [4] and this work: 300 Kwords vs. 500 Mwords, i.e., an order of magnitude of 1000.

Second, adding any of the syntactic features helps, and their combination brings a dramatic improvement over the baseline lexical+POS features: more than +30% in F-score. Furthermore, despite parsing errors of about 15%, syntactic features still improve commas detection with automatic dependency parses by more than 30% absolute.

4.3. Experimental results in Czech

Table 2 compares different feature sets on the Prague Treebank, respectively with manual and automatic parsing.

	Manual deps.		Auto deps.	
	F-sc	SER	F-sc	SER
Lexical	62.9	62.0	62.9	62.0
Lex. + POS	64.1	56.0	**64.1**	56.0
Lex. + POS + IsCoordCz	69.6	49.6	66.6	53.1
Lex. + POS + DepLabel	77.1	39.6	77.0	40.5
Lex. + POS + DepCross	79.6	36.3	78.2	38.1
Lex. + POS + AllSynt	91.2	16.8	**85.5**	27.0

Table 2: Comparison of different feature sets on the Prague Treebank. The confidence interval is ±0.1%.

In Czech, the baseline performances are much higher than in French, as already discussed in section 3.2. The syntactic feature dedicated to handle coordination brings largely significant improvements, about +2.5% absolute. The best feature in Czech is DepCross (+14%), as in French experiments.

Despite parsing errors of about 34%, syntactic features still improve commas detection F-score by +21.4% in absolute value. This clearly confirms the effective importance of syntactic features to recover commas.

5. Discussion

5.1. Feature dependency to the language

Our initial objective in this work was to design generic syntactic features that could be applied to different languages, similarly to the basic word form and part-of-speech tag features, which are applied as is in most natural language processing tasks. This objective has only been partially reached. Indeed, we have observed that the most "basic" syntactic features, such as **DepLabel**, may be very effective in some languages but not on others. We nevertheless proposed such a generic feature, **DepCross**, which seems to work very well in both languages.

Aside from this quest for generic features, we also focused our efforts towards addressing specific usages of commas, such as coordination or modifiers, where syntactic information might intuitively bring valuable information. This approach also gave some improvement in both languages, at the cost of devising and implementing much more complex syntactic features. Nevertheless, such focused features might prove useful, as shown in our experiments, because they address specific patterns that may not be correctly handled by generic features only.

5.2. Robustness to speech recognition errors

Punctuation recovery is a typical language processing task that can be applied to automatic speech transcriptions. One might question the robustness of the proposed syntactic features to speech recognition errors, because of the known limited performances of syntactic parsers on automatic transcriptions. We have not been able so far to test our system on such speech transcriptions, because the French and Prague treebanks are both written text corpora. Furthermore, testing our system on another speech corpus, such as the ESTER corpus, would first require to develop an efficient parser on this type of data. Indeed, domain adaptation of syntactic parsers is known to be extremely difficult, as demonstrated in the CoNLL'2007 campaign, which prevents a direct application of written-text parsers to such corpora. Although we have recently made some progress in this direction [13], there is still no satisfying existing parsing solution nor resources for French spoken data.

6. Conclusions and future work

This work extended previous works dedicated to commas recovery, and in particular [4]. Two new languages are considered, and syntactic features are derived from the dependency tree for each of them. In both cases, the syntactic features improve the performances largely above significance levels. This supports the published conclusions on the importance of syntax for this task and extends them to French and Czech. The next steps will consist in extending this work to support automatic speech recognition outputs, with the objective of enriching such transcripts with punctuation. However, this requires first to solve the weakness of nowadays French and Czech parsers, which are not robust enough to recognition errors.

7. Acknowledgements

This work has been partly supported by the Région Lorraine and the CPER MISN TALC, by a project partly funded by the European Commission. We also thank Mr. Michal Hrala for his help for some experiments.

8. References

[1] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 35–40.

[2] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, 2002, pp. 917–920.

[3] S. M. Shieber and X. Tao, "Comma restoration using constituency information," in *Proc. HLT-NAACL*, 2003, pp. 142–148.

[4] B. Favre, D. Hakkani-Tür, and E. Shriberg, "Syntactically-informed models for comma prediction," in *Proc. ICASSP*, Taipei, Taiwan, April 2009, pp. 4697–4700.

[5] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news," *Speech Communication*, vol. 50, pp. 847–862, 2008.

[6] J. Kolář, J. Švec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech," in *Proc. SPECOM*. Saint-Petersburg: SPIIRAS, 2004, pp. 319–325.

[7] Y. Guo, H. Wang, and J. v. Genabith, "A linguistically inspired statistical model for Chinese punctuation generation," *ACM Transactions on Asian Language Information Processing*, vol. 9, no. 2, p. 27, 2010.

[8] M. Simard, "étude de la distribution de la virgule dans les phrases de textes argumentatifs d'expression française," Ph.D. dissertation, Univ. du Québec Chicoutimi, Apr. 1993.

[9] M.-H. Candito, B. Crabbé, and P. Denis, "Statistical French dependency parsing: treebank conversion and first results," in *Proc. LREC*, La Valletta, Malta, 2010.

[10] J. Hajič, A. Böhmová, E. Hajičová, and B. Vidová-Hladká, "The Prague dependency treebank: A three-level annotation scenario," in *Treebanks: Building and Using Parsed Corpora*, A. Abeillé, Ed. Amsterdam:Kluwer, 2000, pp. 103–127.

[11] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proc. Association of Computational Linguistics*, 2005, pp. 363–370.

[12] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. DARPA Broadcast News Workshop*, Herndon, VA, 1999.

[13] C. Cerisara and C. Gardent, "Analyse syntaxique du français parlé," in *Journée thématique ATALA : Quels analyseurs syntaxiques pour le français ?*, 2009.

Named Entities as New Features for Czech Document Classification

Pavel Král

Dept. of Computer Science & Engineering, and
NTIS - New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
pkral@kiv.zcu.cz

Abstract. This paper is focused on automatic document classification. The results will be used to develop a real application for the Czech News Agency. The main goal of this work is to propose new features based on the Named Entities (NEs) for this task. Five different approaches to employ NEs are suggested and evaluated on a Czech newspaper corpus. We show that these features do not improve significantly the score over the baseline word-based features. The classification error rate improvement is only about 0.42% when the best approach is used.

1 Introduction

Nowadays, the amount of electronic text documents and the size of the World Wide Web are extremely rapidly growing. Therefore, automatic document classification is particularly important for information organization, storage and retrieval.

This work is focused on a real application of the document classification for the Czech News Agency (CTK).[1] CTK produces daily about one thousand text documents, which belong to different classes such as sport, culture, business, etc. In the current application, documents are manually annotated. Unfortunately, the manual annotation represents a very time consuming and expensive task. Moreover, this annotation is often not sufficiently accurate. It is thus beneficial to propose and implement an automatic document classification system.

Named Entity (NE) Recognition was identified as a main research topic for automatic information retrieval around 1996 [1]. The objective is identification of expressions with special meaning such as person names, organizations, times, monetary values, etc. The named entities can be successfully used in many fields and applications, e.g. question answering, information filtering, etc.

In this paper, we propose new features for document classification of the Czech newspaper documents based on the named entities. We believe that NEs bring some additional information, which can improve the performance of our document classification system. Our assumptions are supported by the following observations.

[1] http://www.ctk.eu

A. Gelbukh (Ed.): CICLing 2014, Part II, LNCS 8404, pp. 417–427, 2014.

- NEs can be used to differentiate some similar words according to the context. For example, "Bush" can be American president or popular British band. Using information about the NE, the documents can be classified correctly to one of the two different categories: politics or culture.
- It is possible to use named entities to discover synonyms, e.g. "USA" and "United States" are two different words. However, the word-sense is similar and they represent the same NE, the *country*. This additional information should help to classify two different documents containing "USA" and "United States" words, respectively, into the same category.
- Named entities shall be also used to identify and connect individual words in the multiple-words entities to one token. For example, the words in the expression "Mladá fronta dnes" (*name of the Czech newspaper*) do not have any sense. They can, used separately, produce a mismatch in document classification because the word "dnes" (*today*) is mostly used in the class weather. Using one token can avoid this issue.

Five different approaches to employ this information are proposed and evaluated next.

1. add directly the named entities to the feature vector (which is composed of words (or lemmas)) as new tokens
2. concatenate words related to multiple-word entities to one individual token
3. combine (1) and (2)
4. concatenate words and named entities to one individual token
5. replace words related to the named entities by their NEs

Note that, to the best of our knowledge, named entities were never used previously as features for the document classification task of the Czech documents. Moreover, we have not found another work which uses NEs similarly for document classification.

Section 2 presents a short review about the document classification approaches with the particular focus on the use of the NE recognition in this field. Section 3 describes our approaches of the integration of named entities to the feature vector. Section 4 deals with the realized experiments on the CTK corpus. We also discuss the obtained results. In the last section, we conclude the research results and propose some future research directions.

2 Related Work

Document clustering is an unsupervised approach that aims at automatically grouping raw documents into clusters based on their words similarity, while document classification relies on supervised methods that exploit a manually annoted training corpus to train a classifier, which in turn identifies the class of new unlabeled documents. Mixed approaches have also been proposed, such as semi-supervised approaches, which augment labeled training corpus with unlabeled data [2], or methods that exploit partial labels to discover latent topics [3]. This work focuses on document classification based on the Vector Space Model (VSM), which basically represents each document with a vector of all occurring words weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

Several classification algorithms have been successfully applied [4, 5], e.g. Bayesian classifiers, decision trees, k-Nearest Neighbour (kNN), rule learning algorithms, neural networks, fuzzy logic based algorithms, Maximum Entropy (ME) and Support Vector Machines (SVMs). However, the main issue of this task is that the feature space in VSM is highly dimensional which negatively affects the performance of the classifiers.

Numerous feature selection/reduction approaches have been proposed [6] in order to solve this problem. The successfully used feature selection approaches include Document Frequency (DF), Mutual Information (MI), Information Gain (IG), Chi-square test or Gallavotti, Sebastiani & Simi metric [7, 8]. Furthermore, a better document representation may lead to decreasing the feature vector dimension, e.g. using lemmatization or stemming [9]. More recently, advanced techniques based on Labeled Latent Dirichlet Allocation (LDA) [10] or Principal Component Analysis (PCA) [11] incorporating semantic concepts [12] have been introduced.

Multi-label document classification [13, 14][2] becomes a popular research field, because it corresponds usually better to the needs of the real applications than one class document classification. Several methods have been proposed as presented for instance in surveys [15, 16].

The most of the proposed approaches is focused on English and is usually evaluated on the Reuters,[3] TREC[4] or OHSUMED[5] databases.

Only little work is focused on the document classification in other languages. Yaoyong et al. investigate in [17] learning algorithms for cross-language document classification and evaluate them on the Japanese-English NTCIR-3 patent retrieval test collection.[6] Olsson presents in [18] a Czech-English cross-language classification on the MALACH[7] data set. Wu et al. deals in [19] with a bilingual topic aspect classification of English and Chinese news articles from the Topic Detection and Tracking (TDT)[8] collection.

Unfortunatelly, only few work about the classification of the Czech documents exits. Hrala et al. proposes in [20] a precise representation of Czech documents (lemmatization and Part-Of-Speech (POS) tagging included) and shown that mutual information is the most accurate feature selection method which gives with the maximum entropy or support vector machines classifiers the best results in the single-label Czech document

[2] One document is usually labeled with more than one label from a predefined set of labels.

[3] `http://www.daviddlewis.com/resources/testcollections/reuters21578`

[4] `http://trec.nist.gov/data.html`

[5] `http://davis.wpi.edu/xmdv/datasets/ohsumed.html`

[6] `http://research.nii.ac.jp/ntcir/permission/perm-en.html`

[7] `http://www.clsp.jhu.edu/research/malach/`

[8] `http://www.itl.nist.gov/iad/mig//tests/tdt/`

classification task[9]. It was further shown [21] that the approach proposed by Zhu et al. in [22] is the most effective one for multi-label classification of the Czech documents.

To the best of our knowledge, only little work on the use of the NEs for document classification has been done. Therefore, we will focus on the use of the named entities in the closely related tasks. Joint learning of named entities and document topics has mainly been addressed so far in different tasks than document clustering. For instance, the authors of [23] exploit both topics and named entity models for language model adaptation in speech recognition, or [24] for new event detection. Topic models are also used to improve named entity recognition systems in a number of works, including [25–27], which is the inverse task to our proposed work. Joint entity-topic models have also been proposed in the context of unsupervised learning, such as in [28] and [29].

The lack of related works that exploit named entity recognition to help document classification is mainly explained in [30], which has precisely studied the impact of several NLP-derived features, including named entity recognition, for text classification, and concluded negatively. Despite this very important study, we somehow temper this conclusion and show that our intuition that suggests us that named entity features cannot be irrelevant in the context of document classification, might not be completely wrong. Indeed, nowadays NLP tools have improved and may provide richer linguistic features, and the authors of [30] only use a restricted definition of named entities, which are limited to proper nouns, while we are exploiting more complex types of named entities.

3 Document Classification with Named Entities

3.1 Preprocessing, Feature Selection and Classification

The authors of [20] have shown that morphological analysis including lemmatization and POS tagging with combination of the MI feature selection method significantly improve the document classification accuracy. Therefore, we have used the same preprocessing in our work.

Lemmatization is used in order to decrease the feature number by replacing a particular word form by its *lemma* (base form) without any negative impact to the classification score. The words that should not contribute to classification are further filter out from the feature vector according to their POS tags. The words with approximately uniform distribution among all document classes are removed from the feature vector. Therefore, only the words having the POS tags noun, adjective or adverb remain in the feature vector.

Note that the above described steps are very important, because irrelevant and redundant features can degrade the classification accuracy and the algorithm speed.

In this work, we would like to evaluate the importance of new features. Absolute value of the recognition accuracy thus does not play a crucial role. Therefore, we have chosen the simple Naive Bayes classifier which has usually an inferior classification score. However, it will be sufficient for our experiments to show whether new features bring any supplementary information.

[9] One document is assigned exactly to one label from a predefined set of labels.

3.2 Named Entity Integration

For better understanding, the features obtained by the proposed approaches will be demonstrated on the Czech simple sentence "Český prezident Miloš Zeman dnes navštívil Spojené státy." (*The Czech president Miloš Zeman visited today the United States*) (see Table 1). The baseline features after lemmatization and POS-tag filtration are shown in the first line of this table. The second line corresponds to the English translation and the third line illustrates the recognized named entities.

Table 1. Examples of the NE-based features obtained by the five proposed approaches

	Český	Prezident	Miloš	Zeman	dnes	Spojené	státy.			
	Czech	president	Miloš	Zeman	today	United	States			
	O	O	Figure-B	Figure-I	Datetime-B	Country-B	Country-I			
1.	Český	Prezident	Miloš	Zeman	Figure	dnes	Datetime	Spojené	státy	Country
2.	Český	Prezident	Miloš-Zeman	dnes	Spojené-státy					
3.	Český	Prezident	Miloš-Zeman	Figure	dnes	Datetime	Spojené-státy	Country		
4.	Český	Prezident	Miloš-Zeman-Figure	dnes-Datetime	Spojené-státy-Country					
5.	Český	Prezident	Figure	Datetime	Country					

Named Entities as New Tokens in the Feature Vector - (1). The baseline feature vector is composed of words (lemmas in our case) and their values are calculated by the TF-IDF approach. In this approach, we insert directly the named entity labels to the feature vector as new tokens. The values of the NE features are calculated similarly as the values of the word features using the TF-IDF method. One example of the resulting features of this approach is shown in the first line of the second section of Table 1.

Note that the feature values in all following approaches will be also computed by the TF-IDF weighting.

Concatenation of Words (lemmas) Related to Multiple-word Entities to One Individual Token - (2). As mentioned previously, the individual words of the multiple-word entities have usually the different meaning than connected to one single token. In this approach, all words which create a multiple-word NE are connected together and the NE labels are further discarded. The second line of the second section of Table 1 shows the features created by this approach.

Combination of the Approach (1) and (2) - (3). We assume that the NE labels can bring other information than the connected words of the multiple-word NEs. Therefore, in this approach we combine both previously proposed methods as illustrated in the third line of the second section of Table 1.

Concatenation of Words (lemmas) and Named Entities into One Individual Token - (4). The concatenated words of the multiple-word NEs and their NE labels are used in the previous approach as two separated tokens. In this approach, they are linked together to create one token. This approach should play an important role for word sense disambiguation (e.g. "Bush-Figure" vs. "Bush-Organization"). One example of the features obtained by this approach is shown in the fourth line of the second section of Table 1.

Named Entities Instead of the Corresponding Words - (5). The previously proposed approaches increase the size of the feature vector. In this last approach, the size of the vector is reduced replacing the words corresponding to the named entities by their NE labels. The last line of Table 1 shows the features created by this approach.

Weighting of the Named Entities. We assume that named entities represent the most important words in the documents. Therefore, we further slightly modify the TF-IDF weighting in order to increase the importance of the named entities. The original weight is multiplied by K when named entity identified.

Note that we often use the term "words" in the text while in the experiment we use rather their "lemmas" instead.

4 Experiments

4.1 Tools and Corpora

We used the mate-tools[10] for lemmatization and POS tagging. The lemmatizer and POS tagger were trained on 5853 sentences (94.141 words) randomly taken from the PDT 2.0[11] [31] corpus. The performance of the lemmatizer and POS tagger are evaluated on a different set of 5181 sentences (94.845 words) extracted from the same corpus. The accuracy of the lemmatizer is 81.09%, while the accuracy of our POS tagger is 99.99%. Our tag set contains 10 POS tags as shown in Table 2.

We use the top scoring Czech NER system [32]. It is based on Conditional Random Fields. The overall F-measure on the CoNLL format version of Czech Named Entity Corpus 1.0 (CNEC) is 74.08%, which is the best result so far. We have used the model trained on the private CTK Named Entity Corpus (CTKNEC). The F-measure obtained on this corpus is about 65%.

For implementation of the classifier we used an adapted version of the MinorThird[12] tool. It has been chosen mainly because of our experience with this system.

As already stated, the results of this work shall be used by the CTK. Therefore, for the following experiments we used the Czech text documents provided by the CTK. Table 2 shows the statistical information about the corpus. This corpus is available only for research purposes for free at `http://home.zcu.cz/~pkral/sw/` or upon request to the authors.

In all experiments, we used the five-folds cross validation procedure, where 20% of the corpus is reserved for the test. All experiments are repeated 10 times with randomly reshuffled documents in the corpus. The final result of the experiment is then a mean of

[10] `http://code.google.com/p/mate-tools/`

[11] `http://ufal.mff.cuni.cz/pdt2.0/`

[12] `http://sourceforge.net/apps/trac/minorthird`

Table 2. Corpus statistical information

Unit name	Unit number	Unit name	Unit number
Document	11,955	Numeral	216,986
Category	60	Verb	366,246
Word	5,145,788	Adverb	140,726
Unique word	193,399	Preposition	346,690
Unique lemma	152,462	Conjunction	144,648
Noun	1,243,111	Particle	10,983
Adjective	349,932	Interjection	8
Pronoun	154,232		

Table 3. NE tag-set and distribution in the CTK document corpus

NE	No.	NE	No.	NE	No.	NE	No.
City	55,370	E-subject	5,447	Number	160,633	Religion	24
Country	56,081	Figure	133,317	Organization	119,021	Sport	12,524
Currency	25,429	Geography	7,418	Problematic	17	Sport-club	38,745
Datetime	108,594	Nationality	5,836	Region	14,988	Uknown	1,750

all obtained values. For evaluation of the classification accuracy, we used, as frequently in some other studies, a standard *Error Rate (ER)* metric. The resulting error rate has a confidence interval of $< 0.5\%$.

Our NE tag-set is composed of 16 named entities (see Table 3). This table further shows the numbers of the NE occurrences in the corpus. The total number of the NE occurrences is about 700,000 which represents a significant part of the corpus (approximately 13%).

Note that, this named entities have been identified fully-automatically. Some labeling errors are thus available. This fact can influence the following experiments negatively.

4.2 Analysis of the Named Entity Distribution According to the Document Classes and Classification with Only NEs

This experiment should support our assumption that named entities bring useful information for document classification. Therefore, we realize a statistical study of the distribution of the named entities according to the document classes in the corpus (see Figure 1). This figure shows that some NEs (e.g. E-subject, Region, Sport, etc.) are clearly discriminant across the document classes. The analysis supports our assumption that the NEs can have a positive impact to the document classification.

We further realize another experiment in order to show whether only named entities (without the word features) are useful for the document classification. The results of this experiment (see the first line of Table 4) shows that NEs bring some information for document classification. However, their impact is small.

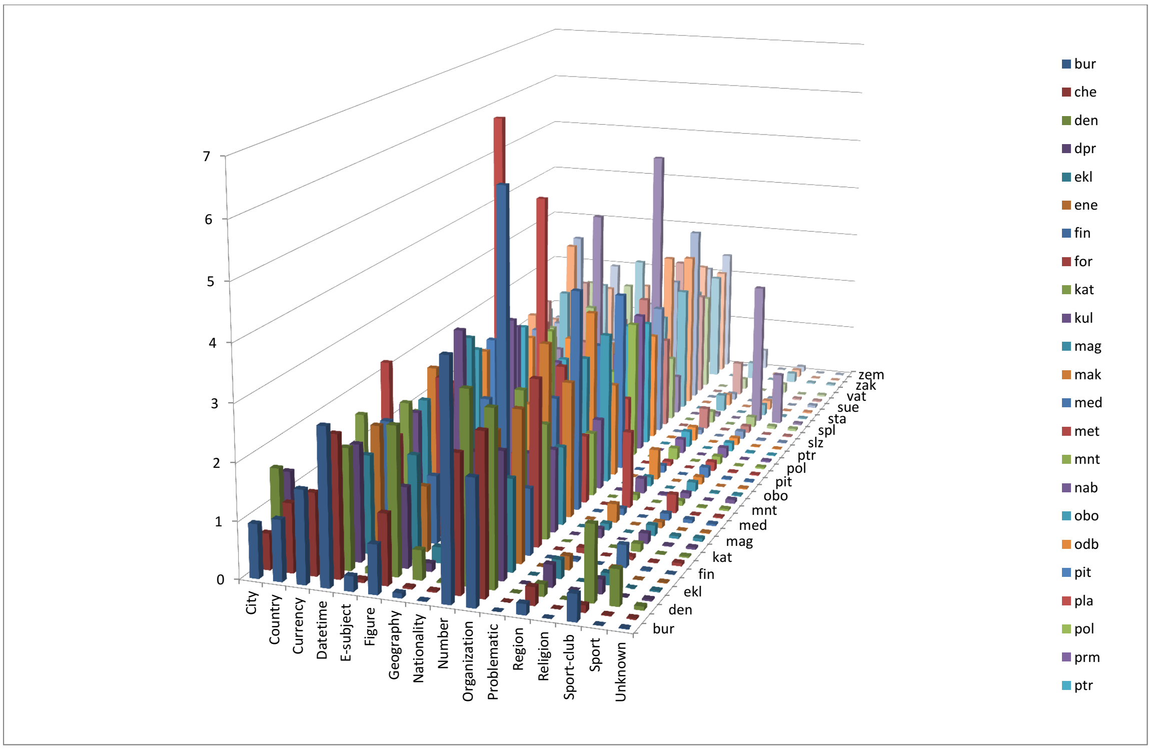


Fig. 1. Distribution of the named entities according to the document classes

4.3 Classification Results of the Proposed Approaches

The Table 4 further shows the recognition error rates of the proposed approaches. We evaluate the NE weights $K \in \{1, 2, 3\}$. The greater weight values are not used because the classification scores is decreasing according to this value in all experiments.

This table shows that the named entities help for document classification only slightly and this improvement is unfortunately statistically not significant. The best score is obtained by the second approach when the words are concatenated across the NEs and the information about the NE labels is completely removed from the feature vector.

Table 4. Document classification error rates [in %] of the different implementations of the named entity features (NE weights $K \in \{1, 2, 3\}$)

Approach		NE weights		
		1	2	3
NEs only		84.25		
Lemmas (baseline)		17.83		
1.	Lemmas + NEs	17.60	17.75	17.79
2.	Concatenated lemmas	**17.41**	18.01	18.71
3.	1 + 2 together	17.48	18.02	18.61
4.	Concatenated Lemmas + NEs as one token	17.54	18.20	18.59
5.	NEs instead of the corresponding words	17.81	18.27	19.03

4.4 Analysis of the Confusion Matrices

In this experiment, we analyze the confusion matrices in order to compare the errors when the baseline word-bases features and the proposed features used (see Table 5). This table illustrates the number of errors, number of different errors in absolute value and in %, respectively. It is depicted that about 27% of errors (except the first proposed method) is different. Therefore, this experiment confirms that named entities bring some additional information. Unfortunately, this information is not sufficient to improve significantly the document classification accuracy on the CTK document corpus.

Note that the error number of the baseline approach is 2,131.

Table 5. Analysis of the confusion matrices errors between the baseline and five proposed approaches

Baseline vs. Proposed Approach		Error no.	Diff. err. no.	Diff. err. no [in %]
1.	Lemmas + NEs	2,104	357	16.97
2.	Concatenated lemmas	2,081	577	27.73
3.	1 + 2 together	2,089	581	27.81
4.	Concatenated Lemmas + NEs as one token	2,097	569	27.13
5.	NEs instead of the corresponding words	2,129	594	27.9

5 Conclusions and Future Work

In this paper, we have proposed new features for the document classification based on the named entities. We have introduced five different approaches to employ NEs in order to improve the document classification accuracy. We have evaluated these methods on the Czech CTK corpus of the newspaper text documents. The experimental results have shown that these features do not improve significantly the score over the baseline word-based features. The improvement of the classification error rate was only about 0.42% when the best approach is used. We have further analyzed and compared the confusion matrices of the baseline approach with our proposed methods. This analysis has shown that named entities bring some additional information for document classification. Unfortunately, this information is not sufficient to improve significantly the document classification accuracy.

However, we assume that this information could play more important role on smaller corpora with more unknown words in the testing part of the corpus. The first perspective thus consists in evaluation of the proposed features on the other (smaller) corpora including more European languages. Then, we would like to propose other sophisticated features which introduce the semantic similarity of word-based features. These features should be useful for example for word-sense disambiguation and can be created for instance by the semantic spaces.

Acknowledgements. This work has been partly supported by the European Regional Development Fund (ERDF), project “NTIS - New Technologies for Information Society”, European Centre of Excellence, CZ.1.05/1.1.00/02.0090. We also would like to thank Czech New Agency (CTK) for support and for providing the data.

References

1. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th Conference on Computational Linguistics, COLING 1996, Copenhagen, Denmark, vol. 1, pp. 466–471. Association for Computational Linguistics (1996)
2. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Mach. Learn. 39, 103–134 (2000)
3. Ramage, D., Manning, C.D., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, pp. 457–465. ACM, New York (2011)
4. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management, pp. 679–694 (2004)
5. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 380–393 (1997)
6. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, ICML 1997, pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
8. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2000, pp. 59–68. Springer, London (2000)
9. Lim, C.S., Lee, K.J., Kim, G.C.: Multiple sets of features for automatic genre classification of web documents. Information Processing and Management 41, 1263–1276 (2005)
10. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, pp. 248–256. Association for Computational Linguistics, Stroudsburg (2009)
11. Gomez, J.C., Moens, M.F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis 56, 741–751 (2012)
12. Yun, J., Jing, L., Yu, J., Huang, H.: A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications 39, 2035–2046 (2012)
13. Novovičová, J., Somol, P., Haindl, M., Pudil, P.: Conditional mutual information based feature selection for classification task. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 417–426. Springer, Heidelberg (2007)
14. Forman, G., Guyon, I., Elisseeff, A.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
15. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34, 1–47 (2002)
16. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3, 1–13 (2007)
17. Yaoyong, L., Shawe-Taylor, J.: Advanced learning algorithms for cross-language patent retrieval and classification. Information Processing & Management 43, 1183–1199 (2007)
18. Olsson, J.S.: Cross language text classification for malach (2004)
19. Wu, Y., Oard, D.W.: Bilingual topic aspect classification with a few training examples. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 203–210. ACM (2008)
20. Hrala, M., Král, P.: Evaluation of the document classification approaches. In: Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierek, A. (eds.) CORES 2013. AISC, vol. 226, pp. 875–884. Springer, Heidelberg (2013)

21. Hrala, M., Král, P.: Multi-label document classification in czech. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS, vol. 8082, pp. 343–351. Springer, Heidelberg (2013)
22. Zhu, S., Ji, X., Xu, W., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 274–281. ACM (2005)
23. Liu, Y., Liu, F.: Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In: Proceedings ASSP (2008)
24. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 297–304. ACM (2004)
25. Guo, H., Zhu, H., Guo, Z., Zhang, X., Wu, X., Su, Z.: Domain adaptation with latent semantic association for named entity recognition. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2009, pp. 281–289. Association for Computational Linguistics, Stroudsburg (2009)
26. Knopp, J., Frank, A., Riezler, S.: Classification of named entities in a large multilingual resource using the Wikipedia category system. PhD thesis, Masters thesis, University of Heidelberg (2010)
27. Zhang, Z., Cohn, T., Ciravegna, F.: Topic-oriented words as features for named entity recognition. In: Gelbukh, A. (ed.) CICLing 2013, Part I. LNCS, vol. 7816, pp. 304–316. Springer, Heidelberg (2013)
28. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 680–686. ACM, New York (2006)
29. Vosecky, J., Jiang, D., Leung, K.W.T., Ng, W.: Dynamic multi-faceted topic discovery in twitter. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, CIKM 2013, pp. 879–884. ACM, New York (2013)
30. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 181–196. Springer, Heidelberg (2004)
31. Hajič, J., Böhmová, A., Hajičová, E., Vidová-Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Abeillé, A. (ed.) Treebanks: Building and Using Parsed Corpora, pp. 103–127. Kluwer, Amsterdam (2000)
32. Konkol, M., Konopík, M.: CRF-based czech named entity recognizer and consolidation of czech NER research. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS, vol. 8082, pp. 153–160. Springer, Heidelberg (2013)

Novel Unsupervised Features for Czech Multi-label Document Classification

Tomáš Brychcín[1,2] and Pavel Král[1,2]

[1] Dept. of Computer Science & Engineering,
Faculty of Applied Sciences,
University of West Bohemia,
Plzeň, Czech Republic
[2] NTIS - New Technologies for the Information Society,
Faculty of Applied Sciences,
University of West Bohemia,
Plzeň, Czech Republic
{brychcin,pkral}@kiv.zcu.cz

Abstract. This paper deals with automatic multi-label document classification in the context of a real application for the Czech News Agency. The main goal of this work consists in proposing novel fully unsupervised features based on an unsupervised stemmer, Latent Dirichlet Allocation and semantic spaces (HAL and COALS). The proposed features are integrated into the document classification task. Another interesting contribution is that these two semantic spaces have never been used in the context of document classification before. The proposed approaches are evaluated on a Czech newspaper corpus. We experimentally show that almost all proposed features significantly improve the document classification score. The corpus is freely available for research purposes.

Keywords: Multi-label Document Classification, LDA, Semantic spaces, HAL, COALS, HPS, Stemming, Czech, Czech News Agency, Maximum Entropy.

1 Introduction

Nowadays, the amount of electronic text documents and the size of the World Wide Web increase extremely rapidly. Therefore, automatic document classification (or categorization) becomes very important for information retrieval.

In this work, we focus on the *multi-label* document classification[1] in the context of a real application for the Czech News Agency (ČTK)[2]. ČTK produces daily about one thousand of text documents. These documents belong to different categories such as politics, sport, culture, business, etc. In the current application, documents are manually annotated. Unfortunately, the manual labeling represents a very time consuming and

[1] *Multi-label* document classification: one document is usually labeled with more than one label from a predefined set of labels vs. *Single-label* document classification: one document is assigned exactly to one label.

[2] http://www.ctk.eu

A. Gelbukh et al. (Eds.): MICAI 2014, Part I, LNAI 8856, pp. 70–79, 2014.

expensive task. It is thus beneficial to propose and implement an automatic document classification system.

One important issue in the document classification field is the high dimensionality and insufficient precision of the feature vector. Several feature selection methods and sophisticated language specific features have been proposed. The main drawback of these methods is that they need a significant amount of the annotated data. Furthermore, a complete re-annotation is necessary when the target language is modified.

In this work, we address these issues by proposing novel fully unsupervised features based on an unsupervised stemmer, Latent Dirichlet Allocation (LDA) and semantic spaces (HAL and COALS). We further integrate these features into the document classification task.

The next scientific contribution is evaluating a new simple LDA model, called S-LDA, which integrates stem features into the topic modeling. Another interesting contribution is the use of semantic space models (i.e. HAL and COALS), because they have not been used for the document classification yet. The last contribution consists in the evaluation of the proposed approaches on Czech, as a representative of morphologically rich language.

The paper structure is as follows. Section 2 introduces the document classification approaches with a particular focus on the document representation. Section 3 describes our proposed features and their integration into the document classification task. Section 4 deals with the experiments on the ČTK corpus. In the last section, we discuss the research results and we propose some future research directions.

2 Related Work

The today's document classification relies usually on supervised machine learning methods that exploit a manually annotated training corpus to train a classifier, which in turn identifies the class of new unlabeled documents. Most approaches are based on the Vector Space Models (VSMs), which mostly represent each document as a vector of all occurring words usually weighted by their Term Frequency-Inverse Document Frequency (TF-IDF).

Several classification algorithms have been successfully applied [3,7], e.g. Bayesian classifiers, decision trees, k-Nearest Neighbor (kNN), rule learning algorithms, neural networks, fuzzy logic based algorithms, Maximum Entropy (ME) and Support Vector Machines (SVMs). However, one important issue of this task is that the feature space in VSM has a high dimension which negatively affects the performance of the classifiers.

Numerous feature selection/reduction approaches have been proposed in order to solve this problem. The successfully used feature selection methods include Document Frequency (DF), Mutual Information (MI), Information Gain (IG), Chi-square test or Gallavotti, Sebastiani & Simi metric [8,9].

In the last years, multi-label document classification becomes a popular research field, because it corresponds usually better to the needs of the real applications than the single-label document classification. One popular approach presented in [27] uses n binary *class/no class* classifiers. A final classification is then given by an union of these partial results. Another approach presented by the authors of [27] simplifies the multi-label document classification task by replacing *each different* set of labels by a new *single label*. Then, a single-label document classifier is created on such data. Note that this approach suffers by the data sparsity problem. Zhu et al. propose in [30] another multi-label document classification approach. The same classifier as in the single-label document classification task is created. The document is associated with a set of labels based on an acceptance *threshold*. The other methods are presented for instance in survey [26].

Furthermore, a better document representation may lead to decreasing the feature vector dimension, e.g. using lexical and syntactic features as shown in [18]. Chandrasekar et al. further show in [6] that it is beneficial to use POS-tag filtration in order to represent a document more accurate. The authors of [21] and [28] use a set of linguistic features. Unfortunately, they do not show any impact to the document classification task. However, they conclude that more complex linguistic features may improve the classification score.

More recently, an advanced technique based on Labeled Latent Dirichlet Allocation (L-LDA) [24] has been introduced. Unlike our approach, L-LDA incorporates supervision by constraining the topic model to use only those topics that correspond to document labels. Principal Component Analysis (PCA) [10] incorporating semantic concepts [29] has been also successfully proposed for the document classification. Semi-supervised approaches, which augment labeled training corpus with unlabeled data [22] were also used.

The most of the proposed approaches is focused on English. Unfortunately, only little work about the document classification in other non-mainstream languages, particularly in Czech, exits. Hrala et al. [14] use lemmatization and POS-tag filtering for a precise representation of the Czech documents. The authors further show the performance of three multi-label classification approaches [13].

3 Document Classification

In the following sections we describe the proposed unsupervised features and classification approaches.

3.1 Unsupervised Stemming

Stemming is a task to replace a particular (inflected) word form by its “stem” (an unique label for all morphological forms of a word). It is used in many Natural Language Processing (NLP) fields (e.g. information retrieval) to reduce the number of parameters with a positive impact to the classification accuracy. Therefore, we assume that stems should improve the results of the document classification.

We propose two approaches to integrate the stem features into the document classification. In the first approach, the stem occurrences are used directly as the features, while in the second one, we use stems as a preprocessing step for LDA. We use an unsupervised stemming algorithm called HPS [5] This stemmer have been already proved to be very efficient in the NLP, see for example [12].

Note that this task is very similar to lemmatization. However, the main advantage of our stemming approach is that it is fully unsupervised and thus it does not need any annotated data (only plain text).

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is a popular topic model that assigns a topic to each word in the document collection. In our first approach, we use a standard LDA model as follows. We calculate the topic probabilities for each document. The probability of each topic t is given by the number of times the topic t occurs in a document divided by the document size. These probabilities are used directly as new features for a classifier.

In our second approach, we use stems instead of words. This concept is motivated by the following assumptions. LDA is a bag-of-word model, thus the word role in a sentence is inhibited. We assume that the morphosyntactic information in a document is useless for inferring topics. Moreover, the word normalization (i.e. stemming in our case) can reduce the data sparsity problem, which is particularly significant in the processing of morphologically rich languages (e.g. Czech). The parameters of such model should be better estimated than the parameters of the standard LDA. The features for the classifier are calculated in the same way as for the word-based LDA. This model will be hereafter called the *S-LDA* (Stem-based LDA).

3.3 Semantic Spaces

Semantic spaces represent words as high dimensional vectors. Semantically close words should be represented by similar vectors and the vector space gives an opportunity to use a clustering method to create word clusters.

The authors of [4] have proved that word clusters created by the semantic spaces improve significantly language modeling. We assume that these models can play an important role for document classification. We use two semantic space models, namely: HAL (Hyperspace Analogue to Language) [19] and COALS (Correlated Occurrence Analogue to Lexical Semantic) [25]. The word clusters are created using Repeated bisection algorithm. The document is then represented as a bag of clusters and we use a tf-idf weighting scheme for each cluster to create the features.

We assume that these models should reduce (analogically as in the previous case) the data sparsity problem. It is worth of mentioning that these two semantic space models have never been used in the context of document classification before.

3.4 Document Classification

For multi-label classification, we use (as presented in [27]) n binary classifiers $C_{i=1}^{n} : d \to l, \neg l$ (i.e. each binary classifier assigns the document d to the label l iff the label is included in the document, $\neg l$ otherwise). The classification result is given by the following equation:

$$C(d) = \cup_{i=1}^{n}: C_i(d) \tag{1}$$

The Maximum Entropy (ME) [1] classifier is used. As a baseline, we use the tf-idf weighting of the word features. Then, this set is progressively extended by the novel unsupervised features. In order to facilitate the reading of the paper, all features are summarized next.

- **Words (baseline)** – Occurrence of a word in a document. Tf-idf weighting is used.
- **Stems** – Occurrence of a stem in a document. Tf-idf weighting is used.
- **LDA** – LDA topic probabilities for a document.
- **S-LDA** – S-LDA topic probabilities for a document.
- **HAL** – Occurrence of a HAL cluster in a document. Tf-idf weighting is used.
- **COALS** – Occurrence of a COALS cluster in a document. Tf-idf weighting is used.

4 Experiments

In our experiments we use LDA implementation from the MALLET [20] tool-kit. For each experiment, we train LDA with 1,000 iterations of the Gibbs sampling. The hyperparameters of the Dirichlet distributions were (as proposed in [11]) initially set to $\alpha = 50/K$, where K is the number of topics and $\beta = 0.1$.

The S-Space package [15] is used for implementation of the HAL and COALS algorithms. For each semantic space, we use a four-word context window (in both directions). HAL uses a matrix consisting of 50,000 columns. COALS uses a matrix with 14,000 columns (as suggested by the authors of the algorithm). SVD (Singular Value Decomposition) was not used in our experiments.

We created the word clusters in the similar way as described in [4], i.e. by using Repeated Bisection algorithm and cosine similarity metric. For clustering, we use an implementation from the CLUTO software package [16]. For both semantic spaces, the word vectors are clustered into four depths: 100, 500, 1,000, and 5,000 clusters.

For multi-label classification we use Brainy [17] implementation of Maximum Entropy classifier.

4.1 Corpus

As mentioned previously, the results of this work will be used by the ČTK. Therefore, we use Czech document collection provided by the ČTK for the training of our models (i.e. LDA, S-LDA, semantic spaces and multi-label classifier).

This corpus contains 2,974,040 words belonging to 11,955 documents annotated from a set of 37 categories. Figure 1 illustrates the distribution of the documents depending on the number of labels. This corpus is freely available for research purposes at `http://home.zcu.cz/~pkral/sw/`.

In all experiments, we use the five-fold cross-validation procedure, where 20% of the corpus is reserved for the test. For evaluation of the document classification accuracy, we use the standard Precision (P), Recall (R) and F-measure (F_m) metrics [23]. The confidence interval of the experimental results is 0.6% at a confidence level of 0.95.

No feature selection has been done in our experiments to clearly show the impact of the proposed features. In the following tables, the term *words* denotes the word features and *stems* denotes the stem features.

4.2 Classification Results of the LDA and S-LDA Models

In this experiment, we would like to compare the classification results of the standalone LDA and S-LDA model (see Table 1). This table shows that the larger number of

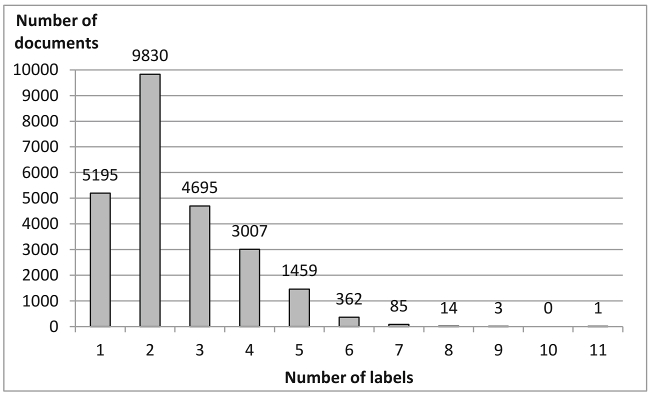


Fig. 1. Distribution of the documents depending on the number of labels

Table 1. Results of stand-alone LDA and S-LDA models

topics	LDA			S-LDA		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	82.9	65.9	73.4	83.1	66.0	73.6
200	83.4	69.1	75.6	83.7	70.3	76.4
300	84.0	71.1	77.0	85.3	72.6	78.4
400	83.3	70.7	77.5	85.5	73.2	78.8
500	84.7	72.6	78.2	85.9	74.0	79.5

the topics is better for document classification. Moreover, the proposed S-LDA slightly outperforms the stand-alone LDA model for all topic numbers.

4.3 Classification Results of the LDA and S-LDA Models with Baseline Word Features

This experiment compares the classification results of the stand-alone LDA and S-LDA models when the baseline word features are also used (see Table 2). Unlike the previous experiment, the recognition score remains almost constant for every topic number and both LDA models. The topic number and LDA type thus no longer play any role for document classification.

Table 2. Results of LDA models with baseline word features

topics	Words+LDA			Words+S-LDA		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	89.0	74.0	80.8	88.9	74.0	80.8
200	88.9	73.8	80.7	88.9	73.6	80.5
300	88.9	73.6	80.6	89.0	73.6	80.6
400	88.8	73.7	80.5	88.8	73.7	80.5
500	88.8	73.7	80.5	88.8	73.5	80.4

4.4 Classification Results of the Semantic Space Models

This experiment compares the classification results of the HAL and COALS models (see Table 3). The table shows that with rising number of clusters the classification score increases. At the level of 5,000 clusters the score is almost the same as for the baseline. However, the number of the parameters in the classifier is significantly reduced.

In the case of COALS and 5,000 clusters the F-measure is slightly better than the baseline. However, we believe this deviation is caused by a chance. In all these experiments COALS outperforms the HAL model.

Table 3. Results of semantic space models

clusters	HAL			COALS		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	58.5	14.7	23.6	66.9	25.2	36.6
500	76.1	51.3	61.3	79.6	59.3	68.0
1000	80.2	62.0	70.0	81.6	64.8	72.2
5000	87.9	72.1	79.2	88.5	73.5	80.3

4.5 Classification Results of the Semantic Space Models with Baseline Word Features

This experiment compares the classification results of the HAL and COALS models when the baseline word features are also used. The results are reported in Table 4. Unlike the previous experiment, the recognition score remains almost constant for all clusters and for both semantic space models.

We can explain this behavior by the fact that the clusters from semantic spaces do not bring any useful additional information compared to the baseline.

Table 4. Results of semantic space models with baseline word features

clusters	Words+HAL			Words+COALS		
	$P[\%]$	$R[\%]$	$F_m[\%]$	$P[\%]$	$R[\%]$	$F_m[\%]$
100	88.2	72.6	79.7	88.2	72.8	79.7
500	88.2	72.7	79.7	88.2	72.7	79.7
1000	88.3	72.8	79.8	88.2	72.7	79.7
5000	88.3	72.8	79.8	88.3	72.7	79.7

4.6 Classification Results of the Different Model Combinations

In this section we evaluate and compare several combinations of our models (see Table 5). The best model configurations from the previous experiments are used. These configurations are compared over the baseline "word" approach (first line in the table). This experiment clearly shows that almost all proposed features significantly improve the document classification accuracy. The F-measure improvement is 2.1% in the absolute value when all proposed features are used. Only the semantic space models do not have any significant impact to improve the classification score. Note that this behavior has been already justified in the previous section.

Table 5. Results of different model combinations. The term COALS denotes the combination of all four COALS models (i.e. 100, 500, 1000, and 5000 clusters). The term HAL denotes the combination of all HAL models. The term S-LDA means the combination of the S-LDA models with 100 and 400 topics.

model	$P[\%]$	$R[\%]$	$F_m[\%]$	impr. $F_m[\%]$
words	88.1	72.7	79.7	
stems	86.4	75.0	80.3	+0.7
words+stems	88.3	74.8	81.0	+1.3
words+HAL	88.4	72.8	79.9	+0.2
words+COALS	88.5	72.8	79.9	+0.2
words+S-LDA	89.2	74.6	81.2	+1.6
words+stems+S-LDA	88.8	75.5	81.6	+1.9
words+stems+S-LDA+COALS	89.0	75.6	81.7	+2.1

5 Conclusions and Future Work

In this work, we have proposed novel fully unsupervised features based on an unsupervised stemmer HPS, Latent Dirichlet Allocation and semantic spaces (HAL and COALS). These features were further integrated into the multi-label document classification task.

We have evaluated the proposed approaches on the ČTK corpus in Czech that is a representative of morphologically rich languages.

We have experimentally shown that almost all proposed unsupervised features significantly improve the document classification score. The F-measure improvement over the baseline is 2.1% absolute, when all proposed features are used.

We plan to extend our work by experiments with different languages and language families. Due to the unsupervised character of the proposed methods, no additional annotations are required.

Acknowledgements. This work has been partly supported by the UWB grant SGS-2013-029 Advanced Computer and Information Systems and by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090. We also would like to thank Czech New Agency (ČTK) for support and for providing the data.

References

1. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics 22(1), 39–71 (1996)
2. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 2003 (2003)
3. Bratko, A., Filipič, B.: Exploiting structural information for semi-structured document categorization. In: Information Processing and Management, pp. 679–694 (2004)
4. Brychcín, T., Konopík, M.: Semantic spaces for improving language modeling. Computer Speech & Language 28(1), 192 (2014)

5. Brychcín, T., Konopík, M.: Hps: High precision stemmer. Information Processing & Management 51(1), 68–91 (2015), http://www.sciencedirect.com/science/article/pii/S0306457314000843
6. Chandrasekar, R., Srinivas, B.: Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging (1996)
7. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(4), 380–393 (1997), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=588021
8. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)
9. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 59–68. Springer, Heidelberg (2000), http://dl.acm.org/citation.cfm?id=646633.699638
10. Gomez, J.C., Moens, M.-F.: Pca document reconstruction for email classification. Computer Statistics and Data Analysis 56(3), 741–751 (2012)
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(Suppl. 1), 5228–5235 (2004)
12. Habernal, I., Ptáček, T., Steinberger, J.: Sentiment analysis in czech social media using supervised machine learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 65–74. Association for Computational Linguistics, Atlanta (2013)
13. Hrala, M., Král, P.: Multi-label document classification in czech. In: Habernal, I., Matousek, V. (eds.) TSD 2013. LNCS, vol. 8082, pp. 343–351. Springer, Heidelberg (2013)
14. Hrala, M., Král, P.: Evaluation of the document classification approaches. In: Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierek, A. (eds.) CORES 2013. Advances in Intelligent Systems and Computing, vol. 226, pp. 875–884. Springer, Heidelberg (2013)
15. Jurgens, D., Stevens, K.: The s-space package: An open source package for word space models. System Papers of the Association of Computational Linguistics (2010)
16. Karypis, G.: Cluto - a clustering toolkit (2003), www.cs.umn.edu/~karypis/cluto
17. Konkol, M.: Brainy: A machine learning library. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2014, Part II. LNCS, vol. 8468, pp. 490–499. Springer, Heidelberg (2014)
18. Lim, C.S., Lee, K.J., Kim, G.C.: Multiple sets of features for automatic genre classification of web documents. Information Processing and Management 41(5), 1263–1276 (2005), http://www.sciencedirect.com/science/article/pii/S0306457304000676
19. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods Instruments and Computers 28(2), 203–208 (1996)
20. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
21. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 181–196. Springer, Heidelberg (2004), http://dx.doi.org/10.1007/978-3-540-24752-4_14
22. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Mach. Learn. 39(2-3), 103–134 (2000), http://dx.doi.org/10.1023/A:1007692713085
23. Powers, D.: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies 2(1), 37–63 (2011)

24. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, vol. 1, pp. 248–256. Association for Computational Linguistics, Stroudsburg (2009), http://dl.acm.org/citation.cfm?id=1699510.1699543
25. Rohde, D.L.T., Gonnerman, L.M., Plaut, D.C.: An improved method for deriving word meaning from lexical co-occurrence. Cognitive Psychology 7, 573–605 (2004)
26. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
27. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3(3), 1–13 (2007)
28. Wong, A.K., Lee, J.W., Yeung, D.S.: Using complex linguistic features in context-sensitive text classification techniques. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 5, pp. 3183–3188. IEEE (2005)
29. Yun, J., Jing, L., Yu, J., Huang, H.: A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications 39(2), 2035–2046 (2012)
30. Zhu, S., Ji, X., Xu, W., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 274–281. ACM (2005)