EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 3, 117–134

MACHINE LEARNING TECHNIQUES IN DIALOGUE ACT RECOGNITION

Mark Fišel

Abstract. This report addresses dialogue acts, their existing applications and techniques of automatically recognizing them, in Estonia as well as elsewhere. Three main applications are described: in dialogue systems to determine the intention of the speaker, in dialogue systems with machine translation to resolve ambiguities in the possible translation variants and in speech recognition to reduce word recognition error rate.

Several recognition techniques are described on the surface level: how they work and how they are trained. A summary of the corresponding representation methods is provided for each technique. The paper also includes examples of applying the techniques to dialogue act recognition.

The author comes to the conclusion that using the current evaluation metric it is impossible to compare dialogue act recognition techniques when these are applied to different dialogue act tag sets. Dialogue acts remain an open research area, with space and need for developing new recognition techniques and methods of evaluation.*

Keywords: conversation analysis, computational linguistics, dialogue act, machine learning, Bayes classifier, hidden Markov model, neural network, decision tree

1. Introduction

Dialogue acts (DAs) have derived from speech acts introduced by Austin (Traum 1999). He introduced the notion of utterances as actions which change the state of the environment, dialogue participants, etc. (in contrast to the notion of utterances as expressions which can be evaluated to true or false). Austin distinguished three types of actions performed by utterances: locutionary (action of saying something –

* This work was supported by the Estonian Science Foundation, grant no. 5685.

shaping the utterance, pronouncing it and using it to refer to real world objects), perlocutionary (action performed by saying something – achieved effects, special to the particular situation – e.g. persuading or surprising) and illocutionary (action performed in saying something – e.g. informing, requesting, asking, answering, warning, apologizing etc.). The latter type has later on been mostly worked with in subsequent research.

Searle extends Austin's work on illocutionary acts (Traum 1999). His main contribution was the attempt to define necessary and sufficient conditions for the act to be performed. These were presented as game definition rules: conditions of normal input/output (necessary for one to express himself and others to understand), propositional content (content restrictions), environment, sincerity (alignment of actual attitudes with the ones expressed in the utterance), etc. Searle also presented several dimensions along which the speech acts can vary, and proposed a speech act taxonomy based on the dimensions.

Dialogue acts were first introduced by Bunt (1994). Although his definition almost repeats the definition of speech acts (DAs – functional units used by the speaker to change the context), the exact notion behind DAs changes from author to author. DAs are also called dialogue moves, utterance types, utterance classes and (returning to the origin) speech acts. DA recognition has also several names, such as recognition, classification, tagging, etc.

Applications. The most common application of automatically recognized dialogue acts are dialogue systems. DAs are used to recognize the intention of the speaker, which helps to determine the necessary response dialogue act (Prasad, Walker 2002), (Fernandez et al. 2005), or even works as shallow parsing – i.e. recognizing DAs is equivalent to understanding the utterance on a more general level (Lendvai et al. 2003).

Another frequent usage of DAs is machine translation in dialogue systems. A correctly recognized DA can help resolve ambiguities in translating utterances. The grammatical form of an utterance doesn't always coincide with the meant intention (e.g. "could you close the window" is not a question, but a request). In addition, different languages have grammatically different polite forms meaning the same thing.

A typical example is the VerbMobil project (Reithinger, Maier 1995, Reithinger et al. 1996, Küssner 1997). There are also other examples (Lee et al. 1997, Levin et al. 2003, etc.).

DAs can be recognized from prosodic and other speech wave features. This can be used to reduce the word recognition error rate in speech recognition (Wright 1998, Wright et al. 1999, Grau et al. 2004, Alshawi 2003).

Hagen and Popowich (2000) describe a dialogue system which is based on a grammar of dialogue acts, which is used to determine the behavior of the whole system and also constrains the possibilities in DA recognition.

Another application of automatic DA recognition is conversational analysis: for instance, it enables selecting subcorpora with specific DA tags, which might be of interest to the researcher.

In Estonia research on DA recognition is done mostly in the University of Tartu. The application is the first one mentioned: recognized DAs are used for determining the speaker's intention and for choosing the proper response DA.

Paper structure. The paper is structured as follows. The 2nd section is devoted to DA taxonomies. It describes the ones used in the projects referred to in this paper. The 3rd section reviews several DA recognition techniques, giving a brief introduction to the technique and describing how it is commonly applied to DA recognition. The 4th section gives an overview of research in DA recognition done in Estonia. The 5th section concludes the paper with a discussion.

2. Dialogue act taxonomies

Although this report is mainly devoted to DA recognition techniques, it is necessary to introduce the main DA taxonomies (or DA tag sets) that are used in the projects, serving as examples for DA recognition.

A DA taxonomy must compromise between two factors. First, the definitions of DA tags must be clear enough in order to be easily separable. If they are not, agreement between human taggers[1] will be low. On the other hand it is efficient to define a reusable taxonomy, which is general enough to be applicable to many different problems.

There seems to be little agreement on how exactly to achieve the compromise. As it can be seen from the projects referred to in this report, many of them prefer using self-defined DA taxonomies. Others use one or another existing taxonomy. The most popular taxonomy, initially designed to be universal, is DAMSL. Other taxonomies have been developed for some corpora, like CallHome or VerbMobil, and have later on gained popularity.

The acronym DAMSL stands for Dialogue Act Markup in Several Layers (Allen, Core 1997). The dialogues are annotated on four different levels, which are the communicative status, the information level, forward-looking function and backward-looking function. It is important to note that some utterances might lack a tag on some levels; for instance an utterance can have the backward-looking function specified, and the forward-looking function missing.

The communicative status isn't marked for most of the utterances. It specifies whether the utterance was uninterpretable, abandoned or was a self talk. The latter indicates that the speaker is not intending the information he talks about for other dialogue participants.

The information level annotation provides an abstract characterization of the content of the utterance. It includes four categories: task fulfilling, task management, communication management and "other level". Task fulfilling means that the utterance is directed at fulfilling the general task of the dialogue, like asking for the time of a flight. Task management indicates an attempt of coordinating the activity of the two speakers; for instance a proposal of switching the current problem or topic. Communication management represents conventional phrases that maintain contact, perception and understanding. These include greetings, closings, acknowledgements, stalling for time (e.g. "let me see"), speech repairs and misunderstandings. The other level category indicates utterances that do not fit into the first three categories: like jokes or small talk.

The forward-looking function corresponds to Austin's illocutionary act, specifying which action the utterance performs. This level includes such categories as

[1] The agreement is most commonly measured with the kappa-statistic (Siegel and Castellan 1988), which equals the percent of utterances which were assigned the same DA tag by all human taggers.

statement, exclamation, information request, etc. The backward looking function level describes the responsive aspect of the utterances, for instance acceptance, rejecting, misunderstanding, full answer, etc.

Another widely used taxonomy based on DAMSL, was designed for the Switchboard corpus (Godfrey et al. 1992) and is called SWBD-DAMSL (Jurafsky et al. 1997). 80% of the tags are present in both the original and the modified version, some tags are added (for instance "non-verbal" category in the communicative status) and some DAMSL categories are subdivided (for instance "statement-non-opinion" and "statement-opinion").

The taxonomy designed for the Spanish dialogue corpus CallHome (Levin et al. 1999) includes in its original version 232 DA tags. These represent the combination of a general category (like statement/question) and the more specific descriptions, e.g. whether the utterance describes the emotional state of the speaker. The original set is frequently collapsed by uniting similar DA tags into one. One of such collapsed variants is the CallHome37, which has all the statements and the back-channels collapsed into a single category, and includes 37 tags. In its turn CallHome37 is collapsed into CallHome10, which includes 8 most general categories (statement, question, answer), a tag for abandoned sentences and a tag for noise.

Among the DA taxonomies designed for single projects, there is the VerbMobil taxonomy (Reithinger, Maier 1995). It also has a single layer of annotation and includes 33 DA tags. These describe the utterance from the point of the performed action; for example *GREET, GIVE REASON, REJECT, FEEDBACK POSITIVE*, etc.

Another taxonomy designed for a single project is the one of the Estonian Dialogue Corpus (EDiC). The taxonomy has two layers of annotation. The upper layer specifies the general typology of the utterance: ritual, questions/answers, directive, additional information, repair, etc. There are twelve types in total, out of which 7 have paired tags (one for initiation and one for response) and 5 have a single tag. The lower layer describes the utterance in greater detail. Each general type has several subtypes on this level: for instance rituals have greeting, thanking, apologizing, etc.; questions/answers have wh-questions, open and closed yes/no questions, refusal to answer, yes/no answers; and so on. In total there are 126 DA tags on the level of detailed annotation. The EDiC DA taxonomy as well as the corpus itself are comprehensively described in (Gerassimenko et al. 2004).

It can be seen that taxonomies vary greatly in size and design, which produces a number of advantages and disadvantages for automatic DA recognition.

3. DA recognition techniques

This chapter describes several techniques successfully applied to DA recognition. The techniques are rarely applied to the textual or speech wave representation of the utterance, because the connection between these and the utterance dialogue act is too complicated for the contemporary techniques to grasp. Therefore both the utterance and the dialogue act have to be encoded in order to be used as the technique's input and output respectively.

A common way to encode an utterance is to describe it and/or its words with features. By content these can be linguistic (sentence structure class, word

morphology, parts-of-speech, etc.), prosodic (intonation changes, speech melody contour classes, etc.), keyword-based (wh-word, some content words, presence binary indicators, etc.), statistical (WEBSOM vectors significance vectors – see sct. 4) etc. The three most common feature data types are binary, numerical and nominal. Numerical features are usually encoded as scalars. Binary true/false values are replaced with 1/0 and also encoded as scalar values. Nominal features are most commonly encoded as vectors, composed of binary indicators, each corresponding to a possible feature value. Thus the component, corresponding to the current value, equals 1, and the other components equal 0. Since scalars can be viewed as 1-component vectors, the encoded features can be concatenated to compose an input vector. The output (the dialogue act) is encoded as a nominal feature.

The less common encoding methods, specific to the technique, will be described in the following subsections, devoted to the corresponding techniques.

The two most commonly used evaluation methods are the percentage of correctly classified utterances and the confusion matrix (the rows of the latter correspond to the original DA tags and the columns – to the hypotheses of the model; thus each cell shows how many utterances of some DA have been classified as another DA tag). This report will focus on the former technique, since the latter is specific to its DA taxonomy, and two confusion matrices of different taxonomies cannot be compared to each other directly.

3.1. N-gram approach

One of the simplest techniques used in DA recognition is the n-grams. It is based on the assumption that the current dialogue act is explicitly determined by k preceding dialogue acts (the Markov assumption). Therefore the candidate for the n-th dialogue act is chosen by the principle

$$c_n = \underset{c}{\operatorname{argmax}}\, P(c|c_{n-1}, \ldots, c_{n-k+1})$$

The conditional probabilities are extracted from tagged corpora by counting all existing DA sequences. These are simply the number of occurrences of the sequence $(c_{n-k+1}, \ldots, c_n)$ in the training corpora, divided by the number of occurrences of a shorter sequence, $(c_{n-k+1}, \ldots, c_{n-1})$. The most common values for k are 2 and 3 (in which case the technique is called, correspondingly, bigrams and trigrams). Using larger values only makes sense when longer dependencies are known to exist in the data.

Larger values heavily increase the sparse data effect. In order to lessen the effect smoothing is sometimes used. One of the most commonly used smoothing techniques is the deleted interpolation. The idea is to use n-gram probabilities of lower order in case the higher order n-gram is absent. The lower-order n-grams are penalized with smaller weights in order to provide privilege for higher-order n-grams.

The VerbMobil project is the most basic example of using n-grams (including deleted interpolation) for dialogue act recognition (Reithinger, Maier 1995). Another example is (Lee et al. 1997), where dialogue acts are used for resolving ambiguities in translations from Korean. Their method is based on the assumption that the

conditional n-gram probability of the current utterance u_n is approximated by the following product:

$$P(u_n|u_{n-1},\ldots,u_{n-k+1}) \approx P(u_n|c_n)P(c_n|c_{n-1},\ldots,c_{n-k+1})$$

3.2. Hidden Markov Models

In case of Hidden Markov Models (HMMs) a process is modeled as two parallel sequences of states, out of which one is observable and the other one – hidden. The states of the hidden sequence comprise a stochastic FSA, with the transition probabilities specified. The two sequences are aligned, and the hidden states are said to "generate" the observable ones, whereas each observable state can be generated by each hidden state with a preset probability. The last parameter, describing a model instance, is the vector of the initial probabilities of the hidden states.

Several assumptions are made about the nature of the dialogue process, the first one being the Markov assumption. The strongest one is that transition and generation probabilities remain constant throughout the process.

All three parameters that describe the model (the initial, transition and generation probabilities) are estimated from the training corpus. The estimation is commonly done with the forward-backward algorithm; applying the model (i.e. finding the most probable hidden state sequence given the observable states) is done with the Viterbi algorithm. A thorough description of HMMs can be found in (Rabiner, Juang 1986).

When modeling dialogues and DAs with HMMs, utterances (or the features that represent them) are the observable outputs and DA tags are the hidden ones. Wright (1998) uses intonation events (intonation fall/rise/etc.) as the features, achieving the precision of 72%. In (Ries 1999) HMMs are applied unconventionally: the input is composed of lattices of words and segments, taken from a speech recognizer. The Viterbi algorithm is used to derive the probabilities of all possible DAs of the utterance, which are then used as input of a multilayer perceptron (see section 3.5), along with several prosodic features. The best achieved precision is 76%.

3.3. Bayes classifiers

The Bayes classifiers get their name from Bayes theorem, which is used in the derivation of the technique. The main idea is the same as with n-gram technique – to maximize the probability of the DA tag c, but here instead of preceding acts, arbitrary features f_1, ..., f_n are allowed to describe the utterance:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c|f_1,\ldots,f_n)$$

Unlike the n-grams, in this case estimating the probability directly from the training data would require it to be of enormous size to avoid the sparse data effect. Instead, some independency assumptions are made, which make the estimation easier. In the extreme case all features are considered to be independent and the DA tag probability is estimated by

$$P(c) \prod_i P(f_i|c)$$

This case is called the naive Bayes classifier.

It is necessary to note however, that the features are rarely fully independent in the strict sense. Instead of assuming them to be, the dependencies can be specified with a directed acyclic graph. For example, with a graph given on figure 1 the DA tag probability is estimated by

$$P(f_i|c) \cdot P(c|f_j, f_k) \cdot P(f_k|f_j) \cdot P(f_j)$$

This case is referred to as Bayesian networks. The dependency network (or the graph) can be composed manually or automatically. The latter way can be achieved by regarding the statistical dependencies between features in the training corpus and ignoring the ones with dependency strength below some quota.

Grau et al. (2004) use the naive Bayes classifier with the bag-of-words method: the feature set is composed of binary features, each indicating the presence or absence of a specific word. They also use a modified version of the classifier (uniform naive Bayes classifier) which neglects the DA probability *P(c)* in equation 1. They achieve a result of 66% on the DAMSL-switchboard corpus and DA taxonomy. Ivanovic (2005) describes application of the same modification to a subset of the DAMSL tag set (12 tags), which results in 80% precision. Levin et al. (2003) also use the bag-of-words, but they use binary grammatical features instead of word features. They achieve a precision of 51% using the NESPOLE corpus (Costantini et al. 2002).

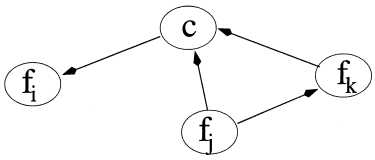

Figure 1. An example of a Bayesian network

Keizer et al. (2002) describe 2 experiments of DA recognition with Bayesian networks. In the first they use three features: sentence type (declarative, imperative, etc., a total of 10 types), subject type (1st/2nd/3rd person) and sentence punctuation (ends with a question or exclamation mark, a full stop, no punctuation, etc.). The network is composed manually and is quite small and simple. They achieve 44% precision with the Schisma corpus (slightly modified DAMSL tag set).

They further apply the technique to their own small dialogue corpus annotated with the same tag set. The used features describe utterances on the surface level: keyword-based features, features indicating whether the sentence starts with some predefined sequence, whether the sentence is/isn't a wh-question, etc.; a total of 13. The network is generated automatically in an iterative manner: first a small part of the corpus was tagged with DAs manually, and then a network was generated based on that part. After training the network had an average precision of 69%. The resulting network was applied to a bigger part of the corpus, and a new network was generated based on that bigger part. Training the new network resulted in 83% precision.

3.5. Multilayer perceptrons

Multilayer perceptrons (MLP) is one of the most frequently used neural network techniques, in general as well as in DA recognition. Neural networks are designed analogically to the human brain. They are capable of learning complex non-linear dependencies between input and output.

A neural network is composed of simple computational units (called neurons), which are organized into a network. The neurons have input and output connections, which are used to connect them. These connections are weighed, which means that the signal is multiplied by the connection weight – e.g. if the weight is 0, the connection doesn't let the signal through, or if it is 1, the signal is strong.

Neurons have output values which are computed based on their inputs, which are other neuron output values, weighed by the incoming connections. The network output complexity is achieved by combining the simple functions, that each neuron implements.

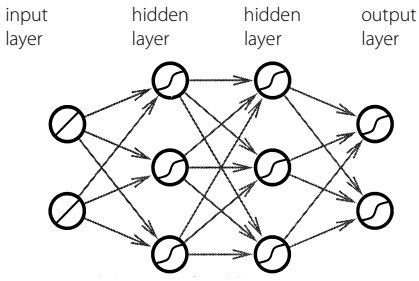

Figure 2. An example of a multilayer perceptron

MLP neurons are grouped into a linearly ordered set of layers, whereas the first layer is assigned the role of the input layer, and the last one – of the output layer. Connections are only allowed in the increasing direction. The function which neurons implement can be different, but must be monotonic and differentiable; it is usually chosen to be a sigmoid. Figure 2 shows a multilayer perceptron with 2 hidden layers.

The most common training algorithm for MLP is error back-propagation. Its description, as well as other algorithms can be found in (Haykin 1999).

Wright (1998) describes training an MLP with one hidden layer on the DCIEM corpus (Bard et al. 1996) (annotated with 12 different DA tags). Using only prosodic features extracted from the speech waves (a total of 54 binary features) she achieves 70% precision.

In Sanchis and Castro (2002) bag-of-words is used to form the MLP input. Before the method is applied all words are replaced with 3 types of categories: words in their basic forms (verbs in infinitive, nouns in singular, adjectives in singular and without gender, etc.), task-specific categories (departure and arrival cities, train types, etc.) and general categories (city names, days of week, months, numbers, etc.). Furthermore the infrequent words are ignored. In total the vocabulary size is 138 categories. The tag set includes 16 tags; the resulting MLP accuracy is 92%.

Levin et al. (2003) apply MLP to the same input as the naive Bayes classifier (binary grammatical feature bag-of-words), achieving the precision of 72% for English and 68% for German.

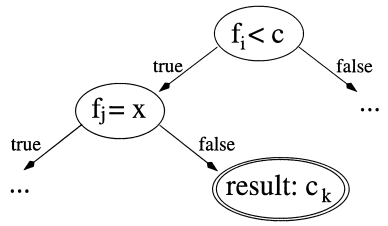

Figure 3. An example of a decision tree

3.6. Decision trees

Decision trees (or classification and regression trees, CART) have the advantage of being a white-box type method, meaning that it is easy for a human to interpret a trained model (relatively to the black-box models, like HMM or MLP).

Each node in a decision tree contains a set of conditions, describing which arc to continue traversing. Each decision is based on one of the parameters, describing the input sample (utterance in case of DA recognition). Each leaf is assigned a DA tag. Recognition is then performed by starting from the tree root and moving along the arcs according to the decisions. When a leaf is reached, the utterance is assigned the DA tag which corresponds to the leaf. A decision tree example can be found on figure 3.

Training decision trees stands for automatically composing them. A common algorithm is the Ross Quinlan ID3 algorithm. The main idea is to select into the tree root the parameter which causes the entropy of the divided subsets to be minimal; the entropy is defined as

$$H = -\sum_i p_i \log_2(p_i)$$

where p_i is the number of utterances, assigned the tag i according to the new tree. The algorithm therefore aims at composing as small a tree as possible.

Wright (1998) applies decision trees to the same data as MLP (binary features extracted from speech waves). Levin et al. (2003) also use the same data as with MLP and naive Bayes classifier (binary grammatical feature bag-of-words). Both methods result in 70% precision.

3.7. Transformation-based learning

Transformation-based learning (TBL) was introduced by Brill (1993). It is based on a set of rules, which are applied consecutively to the data, changing some tags into other ones. The rules are controlled by preset templates; the most common ones are of the type "if current tag is A, it is preceded by tag B and/or the word C is present in one of the preceding N utterances, change the current tag to D".

Rules are composed in a supervised manner. Having a marked training corpus, all possible rules are generated from the templates, after which the rules are selected iteratively: the rule bringing the biggest improvement to the precision is selected on each iteration. The process is continued until one of the stopping criteria is met;

the most common is that no improvement is brought by applying any rule.

Since the total amount of all possible rules can be huge, it is computationally expensive to test each rule, especially since most of them bring precision degradation. One way to improve the situation is to use the Monte-Carlo pruning method. According to that, a fixed number of rules is selected by random; only these rules are later tested. Although this might exclude the very best rule from the selected set, it is highly probable that the set will contain a rule which will still bring a lot of improvement, even if not the maximum.

Samuel et al. (1998) describe applying transformation-based learning with the Monte-Carlo optimization to the VerbMobil corpus and DA tag set. Besides the usual TBL features (i.e. the neighboring utterances and DAs) they use the speaker, punctuation information, word and dialogue act cue[2] presence indicators. They achieve a precision of 75%.

Samuel et al. (1999) describe using dialogue act cues only, also applying TBL to the VerbMobil corpus. Although the focus of the paper is the automatic selection of the DA cues, they describe the DA recognition experiments, which result in a model with 72% precision.

3.8. Memory-based learning

Memory-based learning (MBL) is a relatively simple supervised classification technique. It exploits the idea that handling new data can be done by comparing it to earlier experience, instead of extracting rules from it and applying them. Therefore in MBL the training samples are stored into memory and the new unseen samples are compared to them.

There are several methods of comparing the stored and the unknown new samples. The simplest and the mostly used one is the *k*-nearest-neighbors classification. This method consists of defining a distance measure between the samples, and finding *k* stored samples which have the smallest distance to the new, unclassified sample. It is assumed that the "nearest" samples are similar to the new sample and therefore their classification can be extrapolated on it. The new sample is therefore assigned the class which dominates among the "neighbors".

A common way to store the training samples is to replace them with *n* features, and to save the classification info. The simplest distance metric for this representation is the overlap metric. It simply sums the individual feature distances of the two samples. If a feature is scalar or vector, Euclidean distance is used, normalized with the maximum value for the given feature. In case of nominal features the individual distance is 1 if the values differ and 0 if they coincide. The combination of k-NN classification with the overlap metric is called the IB1 algorithm. Other metrics as well as a comprehensive description of memory-based learning can be found in the Tilburg Memory-Based Learner (TiMBL) reference guide (Daelemans et al. 2004).

Levin et al. (2003) apply TiMBL to the corpora of the NESPOLE machine translation project. They used the IB1 algorithm with 1 neighbor. The features are the same as in their MLP, decision tree and naive Bayes experiments (binary grammatical feature bag-of-words). They achieve a precision of 70%.

[2] Word sequences that are often met in the dialogue.

Lendvai et al. (2003) also use IB1. The features they use include prosodic features extracted from speech waves, bag-of-words vectors based on the speech recognition guesses about the pronounced words and context features (preceding tags etc.). The accuracy of the trained learner is 73.5%.

Fernandez et al. (2005) use features extracted from morphological and PoS information (such as the presence of a wh-word, repeated words in the utterance and its predecessor, etc.). They use the modified value difference metric (Daelemans et al. 2004), claiming that it performs better than the default (overlap) metric. The technique results in 87% precision.

3.9. Less frequently used techniques

Küssner (1997) applies a rule-based approach with the FLEX++ system to DA recognition. The used features include prosodic, semantic, syntactical as well as the DA guess of some unmentioned statistical DA tagger. Since the papers focus is not explicitly DA recognition, no experiment results are reported. In Lendvai et al. (2003) another rule-based approach is used, which is the RIPPER algorithm. The data representation is the same as in their experiment with memory-based learning. The achieved precision is 60%.

Serafin et al. (2003) and Serafin, Eugenio (2004) describe using latent semantic analysis (LSA) for DA recognition. The key-point idea is taking the word-document matrix (each component equals the number of times a word occurs in a document), compressing it with Singular Value Decomposition (SVD) and later comparing compressed new utterances to the matrix using Euclidean distance. The best reported result is 79%.

Tanaka and Yokoo (1999) describe alternative approach where DA tagging is integrated with discourse segmentation. The evaluation method is labeled bracket matching (Nagata 1994), the authors report the highest precision to be 75%.

4. DA recognition research in Estonia

In Estonia systematic research in the area of dialogue systems is done at the University of Tartu. The project incorporating the research aims at building a natural language interface to a database of transport timetables. The application of DA recognition is therefore the one that is common for dialogue systems: to recognize the intention behind the speaker's utterance and generate a proper response.

The peculiarity of DA recognition in this case lies in the high granularity of the DA taxonomy. The relative incidence of several DAs is low, which means both that their definition is very specific and that it's hard for a machine learning technique to not skip them in the learning process.

Fishel (2005) and Fišel, Kikas (2006) describe adapting several text classification preprocessing methods to DA recognition. WEBSOM-style preprocessing (Honkela et al. 1997), where the features describe the whole utterance, is tested with multilayer perceptrons. The same preprocessing method is adapted for processing the utterance one word at a time with simple recurrent networks: this way words are

represented with either random vectors (with each component ranging between –1 and 1) or random binary vectors (where each component is either –1 or 1). A different approach is coding the words with significance vectors (Wermter 2000). Finally, complete morphological description of each word as a single feature is tested.

All the experiments with neural networks were done with the upper layer of the DA taxonomy (see section 2). The best accuracy, showed by recurrent networks with significance vector coding, is 58%; the accuracy of other techniques ranges between 40% and 55%. The networks succeed at classifying DA tags represented in many utterances, but completely ignore the less frequent ones.

Another applied technique are decision trees; the experiments are described in Kikas (2005) and Fišel, Kikas (2006). Several types of input features are used in this case. First of them are the keyword-based features; intonation marks are also exploited. Secondly, the neighboring DAs are also included as a hint to the classifier. Finally the most frequent bi- and trigrams are used. The experiments are focused on the detailed level of the EDiC DA taxonomy; the described task is therefore to distinguish between126 different tags.

The resulting accuracy is 45%. The decision trees appear to be capable of classifying the most frequent DA tags as well as neural networks; they also succeed in recognizing other DAs which are less frequent. However, DA tags represented in less than 1% of utterances are also ignored.

Currently a new algorithm of DA recognition is being developed. The general idea is to find regular expressions that best match utterances of definite DAs. These expressions can then be combined into a single decision tree, where each token of the expression serves as a tree node, which has two child-nodes: one for matching utterances and one for not matching ones. No experimental results have been published so far.

The latest experiments with the EDiC taxonomy detailed level exploit a Bayes-like classifier. The used features are the 2 preceding DA tags (equivalent to trigrams), the number of words *k* and the utterance words. In order to take into consideration the different number of words in every utterance, word probabilites are averaged with a geometrical mean:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c|c_{n-1}, c_{n-2}) \cdot P(c|k) \cdot \left(\prod_{i=1}^{k} P(w_i|c) \right)^{\frac{1}{k}}$$

The resulting accuracy is 61%, which is noticeably close to the EDiC taxonomy kappa-statistic: whereas the statistic was measured on 45 dialogues and equalled 73%, the current results were measured with 10-fold cross-validation on 776 dialogues. However, in order to compare the two directly, the results should be calculated on the same dialogue sets.

As shown by this section, the statistical methods tested so far did not succeed to fully solve the taxonomy. On the other hand, fully describing it with manually composed rules seems not to be feasible given the complexity of the taxonomy considering the available resources. The author believes that further improving the DA recognition for the full EDiC taxonomy would require combining rule-based methods with statistical ones. For instance, specific errors of the Bayes classifier can be corrected by adding manually deduced rules.

5. Discussion

Several DA recognition techniques have been reviewed; a summary of the referred projects can be found in tables 1, 2 and 3 in appendix A.

The reported precisions range from 50% to 95%, whereas in several cases the precisions of applying the same technique to different corpora and/or tag sets differ greatly (e.g. the precision of naive Bayes is in different cases 50%, 66% and 82%). Therefore the technique choice doesn't seem to be the most important one.

It is possible to deduce filigree influences of exact combinations of recognition techniques and representation methods on the precision, but according to the author's opinion it would not make any sense. The hit-count precision does not take into consideration the differences of the DA taxonomies (size, definition clarity, design structure). Although it's a convenient and simple value, it can't be a basis of a conclusion that one technique or data representation is better than the other, unless the used DA taxonomy was the same.

In order to compare DA recognition techniques based on the results of experiments, applying them to different DA taxonomies, an evaluation metric has to be developed, which would take into consideration the DA tag set parameters, possibly also the size of training/testing corpora, training quality etc.

The conclusion is that since none of the techniques appear to be ideal or fully robust, and proper evaluation is absent, dialogue acts are an open research area, requiring a lot of work concerning recognition, representation and evaluation methods.

References

Allen, James; Core, Mark 1997. Draft of DAMSL: Dialog Act Markup in Several Layers, Technical Report, Discourse Research Initiative.

Alshawi, Hiyan 2003. Effective Utterance Classification with Unsupervised Phonotactic Models. – Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1. Edmonton, Canada, 1–7.

Bard, Ellen; Sotillo,Cathy; Anderson, Anne; Taylor, Martin 1996. The DCIEM Map Task Corpus: Spontaneous Dialogue Under Sleep Deprivation and Drug Treatment. – Proceedings of ICSLP'96. Philadelphia, PA, USA, 1958–1961.

Brill, Eric 1993. A Corpus-Based Approach to Language Learning. PhD thesis. Department of Computer and Information Science, University of Pennsylvania.

Bunt, Harry C. 1994. Context and Dialogue Control. – THINK 3, 19–31.

Daelemans, Walter; Zavrel, Jakub; van der Sloot, Ko; van den Bosch, Antal 2004. TiMBL: Tilburg Memory-Based Learner Reference Guide. Technical Report ILK 04-02. Tilburg University and University of Antwerp.

Fernandez, Raquel; Ginzburg, Jonathan; Lappin, Shalom 2005. Using Machine Learning for Non-Sentential Utterance Classification. – Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal, 77–86.

Fishel, Mark 2005. Dialogue Act Recognition in Estonian Dialogues Using Artificial Neural Networks. – Proceedings of the 2nd Baltic Conference on Human Language Technologies. Tallinn, 231–235.

Fišel, Mark; Kikas, Taavet 2006. Dialoogiaktide automaatne tuvastamine. – Mare Koit, Renate Pajusalu, Haldur Õim (toim.). Keel ja arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6. Tartu: Tartu Ülikooli Kirjastus, 233–245.

Gerassimenko, Olga; Hennoste, Tiit; Koit, Mare; Rääbis, Andriela; Strandson, Krista; Valdisoo, Maret; Vutt, Evely 2004. Annotated Dialogue Corpus as a Language Resource: An Experience of Building the Estonian Dialogue Corpus. – Proceedings of the 1st Baltic Conference on Human Language Technologies. Riga, Latvia.

Godfrey, John J.; Holliman, Edward; McDaniel, Jane 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. – Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, CA, USA, 517–520,

Grau, Sergo; Sanchis, Emilio; Castro, Maria Jose; Vilar, David 2004. Dialogue Act Classification Using a Bayesian Approach. – Proceedings of the 9th International Conference Speech and Computer, 495–499.

Hagen, Eli; Popowich, Fred 2000. Flexible Speech Act Based Dialogue Management. – Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue. Hong Kong, 131–140.

Haykin, Simon 1999. Neural Networks: A Comprehensive Foundation. 2nd ed. Prentice-Hall, Inc.

Honkela, Timo; Kaski, Samuel; Lagus, Krista; Kohonen, Teuvo 1997. WEBSOM – Self-Organizing Maps of Document Collections. – Proceedings of Workshop on Self-Organizing Maps. Helsinki, 310–315.

Ivanovic, Edward 2005. Dialogue Act Tagging for Instant Messaging Chat Sessions. – Proceedings of the ACL Student Research Workshop. Ann Arbor, Michigan, 79–84.

Jurafsky, Daniel; Shriberg, Elizabeth; Biasca, Debra 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, draft 13, technical report 97-01. Institute of Cognitive Science, University of Colorado.

Keizer, S.; op den Akker, R.; Nijholt, A. 2002. Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. – 3rd SIGdial Workshop on Discourse and Dialogue. Philadelphia, USA, 88–94.

Kikas, Taavet 2005. Dialoogiaktide tuvastamine eestikeelsetes dialoogides otsustuspuude abil. Bakalaureusetöö. Käsikiri Tartu Ülikooli arvutiteaduse instituudis.

Küssner, Uwe 1997. Applying DL in Automatic Dialogue Interpreting. – Proceedings of the International Workshop on Description Logics. Yvette, France, 54–58.

Lee, Jjae-won; Kim, Gil Chang; Seo, Jungyun 1997. A Dialogue Analysis Model with Statistical Speech Act Processing for Dialogue Machine Translation. – Proceedings of the Spoken Language Translations EACL'97 Workshop. Budapest, Hungary, 10–15.

Lendvai, Piroska; van den Bosch, Antal; Krahmer, Emiel 2003. Machine Learning for Shallow Interpretation of User Utterances in Spoken Dialogue Systems. – Proceedings of the EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management. Budapest, Hungary, 69–78.

Levin, Lori; Langley, Chad; Lavie, Alon; Gates, Donna; Wallace, Dorcas; Peterson, Kay 2003. Domain Specific Speech Acts for Spoken Language Translation. – Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo, Japan.

Levin, Lori; Ries, Klaus; Thyme-Gobbel, Ann; Levie, Alon 1999. Tagging of Speech Acts and Dialogue Games in Spanish Call Home. – Proceedings of the ACL Workshop "Towards Standards and Tools for Discourse Tagging". Somerset, NJ, USA, 42–47.

Nagata, Masaaki 1994. A Stochastic Japanese Morphological Analyzer Using a Forward-DP and Backward-A* N-best search algorithm. – Proceedings of the 15th conference on Computational linguistics, Vol. 1. Kyoto, Japan, 201–207.

Prasad, Rashmi; Walker, Marilyn 2002. Training a Dialogue Act Tagger for Human-Human and Human-Computer Travel Dialogues. – Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue. Philadelphia, Pennsylvania, 162–173.

Rabiner, Lawrence; Juang, Bbiing-Hwang 1986. An Introduction to Hidden Markov Models. – ASSP Magazine, IEEE 3 (1), 4–16.

Reithinger, Norbert; Engel, Ralf; Kipp, Michael; Klesen, Martin 1996. Predicting Dialogue Acts for a Speech-To-Speech Translation System. – Proceedings of the 4th International Conference on Spoken Language Processing, Vol. 2. Philadelphia, Pennsylvania, 654–657.

Reithinger, Norbert.; Maier, Elisabeth 1995. Utilizing Statistical Dialogue Act Processing in VERBMOBIL. – Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, Massachusetts, 116–121.

Ries, Klaus 1999. HMM and Neural Network Based Speech Act Detection. – IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1. Phoenix, Arizona, 497–500.

Samuel, Ken; Carberry, Sandra; Vijay-Shanker, K. 1998. Dialogue Act Tagging with Transformation-Based Learning. – Proceedings of the 17th International Conference on Computational linguistics, Vol. 2. Montreal, Quebeck, Canada, 1150–1156.

Samuel, Ken; Carberry, Sandra; Vijay-Shanker, K. 1999. Automatically Selecting Useful Phrases for Dialogue Act Tagging. – Proceedings of the 4th Conference of the Pacific Association for Computational Linguistics. Waterloo, Ontario, Canada.

Sanchis, Emilio; Castro, Maria Jose 2002. Dialogue Act Connectionist Detection in a Spoken Dialogue System. – Proceedings of the 2nd International Conference on Hybrid Intelligent Systems. Santiago de Chile, Chile, 644–651.

Serafin, Riccardo; Eugenio, Barbara D. 2004. FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification. – Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 692–699.

Serafin, Riccardo; Eugenio, Barbara. D.; Glass, Michael 2003. Latent Semantic Analysis for Dialogue Act Classification. – Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 2. Edmonton, Canada, 94–96.

Siegel, Sidney; Castellan, John N. 1988. Nonparametric Statistics for the Behavioral Sciences. 2nd ed. New York: McGraw-Hill.

Tanaka, Hideki; Yokoo, Akio 1999. An Efficient Statistical Speech Act Type Tagging System for Speech Translation Systems. – Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, MD, USA, 381–388.

Traum, David R. 1999. Speech Acts for Dialogue Agents. – Anand Rao, Michael Wooldridge (Ed.). Foundations and Theories of Rational Agents. Kluwer Academic Publishers, 173–206.

Wermter, Stefan 2000. Neural Network Agents for Learning Semantic Text Classification. – Information Retrieval 3 (2), 87–103.

Wright, Helen 1998. Automatic Utterance Type Detection Using Suprasegmental Features. – Proceedings of the 5th International Conference on Spoken Language Processing, Vol. 4. Sydney, Australia, 1403.

Wright, Helen; Poesio, Massimo; Isard, Stephen 1999. Using High Level Dialogue Information for Dialogue Act Recognition Using Prosodic Features. – Proceedings of an ESCA Tutorial and Research Workshop on Dialogue and Prosody. Eindhoven, The Netherlands, 139–143.

Appendix: Summary of DA Recognition Projects

This appendix includes the tables summarizing the DA recognition methods, data representation methods and results of projects, referred to in the DA recognition techniques section.

Table 1. DA recognition in dialogue systems

Project reference	DA taxonomy	Recognition technique	Data representation	Best result
(Keizer et al. 2002)	DAMSL	Bayes	linguistic, punctuation	44%
(Ivanovic 2005)	SWBD-DAMSL	Bayes	BoW, DA n-grams	82%
(Samuel et al.1999)	VM	TBL	DA cues	72%
(Samuel et al.1998)	VM	TBL with Monte Carlo	speaker, punctuation, DA cues	75%
(Ries 1999)	CallHome	HMM+MLP	prosodic and word features	76%
(Serafin and Eugenio, 2004)	CallHome	FLSA	W*D matrix	79%
(Prasad and Walker, 2002)	DARPA	RIPPER	context and word features	99%
(Fishel and Kikas, 2006)	EDiC	Bayes	words, n-grams	61%
		CART	keywords, n-grams	45%
(Lendvai et al. 2003)	self-defined	TIMBL	prosodic, BoW	74%
		RIPPER	(same)	60%
(Fernandez et al. 2005)	self-defined	SLIPPER	linguistic and antecedent	87%
		TIMBL	(same)	87%
		MaxEnt	(same)	87%
(Sanchis and Castro, 2002)	self-defined	MLP	reduced lexicon BoW	92%

Table 2. DA recognition in machine translated dialogue systems

Project reference	DA taxonomy	Recognition technique	Data representation	Best result
(Küssner 1997)	VM	FLEX++	prosody, syntax, semantics	
(Reithinger and Maier 1995)	VM	n-grams	DA n-grams	81%
(Reithinger et al. 1996)	VM	n-grams	DAs, grammar	76%
(Levin et al. 2003)	NESPOLE	TIMBL	grammatical	70%
		CART	(same)	70%
		MLP	(same)	70%
		Bayes	(same)	50%
(Lee et al. 1997)	self-defined	n-grams	syntax patterns, DA n-grams	79%
(Tanaka and Yokoo 1999)	self-defined	self-defined probabilistic method	morpheme sequences	75%

Table 3. DA recognition in speech recognition

Project reference	DA taxonomy	Recognition technique	Data representation	Best result
(Wright et al. 1999)	DCIEM	CART	prosody features	69%
(Wright 1998)	DCIEM	CART	intonation events	71%
		HMM	prosodic features	72%
		MLP	prosodic features	70%
(Grau et al., 2004)	SWBD-DAMSL	Bayes	n-grams	66%
(Alshawi 2003)	AT&T	HMM	phoneme sequences	95%

Mark Fišel (Tartu Ülikool) on erialalt informaatik. Uurimisvaldkonnad: masinõpe, masintõlge, dialoogisüsteemid.
fishel@ut.ee

MASINÕPPE TEHNIKAD DIALOOGIAKTIDE TUVASTAMISES

Mark Fišel
Tartu Ülikool

Artiklis käsitletakse dialoogiakte, nende rakendusi ja automaatse tuvastamise tehnikaid – nii Eestis kui ka mujal maailmas. Kirjeldatakse kolme põhilist rakendust: dialoogisüsteemides selleks, et tuvastada rääkija kavatsusi, masintõlkega dialoogsüsteemides selleks, et lahendada mitmesust võimalikkude tõlkevariantide vahel, ning kõnetuvastuses selleks, et vähendada sõnade tuvastuse vigade arvu.

Kirjeldatakse mitmeid tuvastamistehnikaid. Kõige sagedamini kasutatavad neist on Markovi peitmudel, mitmekihiline tajur, naiivne Bayesi klassifitseerija, Bayesi võrgud, otsustuspuud, mälupõhine ning transformeerimispõhine õppimine. Iga tehnikat kirjeldatakse üldisel tasemel: töötamise printsiipe ning seda, kuidas toimub treenimine. Artikkel sisaldab mitmeid näiteid selle kohta, kuidas neid tehnikaid saab rakendada dialoogiaktide tuvastamiseks. Eraldi pööratakse tähelepanu dialoogiaktide tuvastamise uurimisele Eestis.

Autor jõuab järeldusele, et tuvastamistehnika valik ei olegi dialoogiaktide puhul kõige olulisem. Mõjuvaim faktor on hoopis dialoogiaktide süsteemi karakteristikud (suurus, aktide definitsioonide selgus jms); samas aga enim kasutatavad hindamiskriteeriumid ei võta neid karakteristikuid arvesse. Seega on võimatu võrrelda tuvastamise meetodeid selliste eksperimentide alusel, kus neid rakendatakse erinevatele dialoogiaktide süsteemidele.

Dialoogiaktid on avatud uurimisala, kus on võimalust ja vajadust arendada uusi tuvastamistehnikaid ning hindamise kriteeriume.

Võtmesõnad: vestlusanalüüs, arvutuslingvistika, dialoogiakt, masinõpe, Bayesi klassifitseerija, Markovi peitmudel, neurovõrk, otsustuspuu