CONFIDENCE MEASURES FOR SEMI-AUTOMATIC LABELING OF DIALOG ACTS

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

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

Jana Klečková[2]

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

ABSTRACT

This paper deals with semi-supervised classifier training for automatic Dialog Acts (DAs) recognition. In our previous works, we have designed a dialog act recognition system for reservation applications in the Czech language. In this work, we propose to retrain this system on another corpus, for another task (broadcast news speech), in a different language (French) and with another set of dialog acts. This is realized using a semi-supervised approach based on the Expectation-Maximization (EM) algorithm. We show that, in the proposed experimental setup, the use of confidence measures to filter out incorrectly recognized dialog acts is required to improve the results. Two confidence measures are thus proposed and evaluated on the French broadcast news corpus. Experimental results confirm the interest of this approach for the task of training automatic dialog act classifiers.

Index Terms— Confidence measure, expectation maximization, dialog act, semi-supervised training

1. INTRODUCTION

This work deals with automatic recognition of dialog acts, such as questions, statements, agreements, backchannels, etc. The system developed was previously evaluated on a Czech corpus, with a limited set composed of four DAs [1, 2]. In this work, we validate our approach on another corpus, another language (French) and with a larger set of DAs.

One of the main issue in the domain of automatic dialog acts recognition concerns the lack of training data and the design of a fast and cheap method to label new corpora. We propose to apply the general semi-supervised training approach based on the Expectation-Maximization algorithm to the task of labeling a new corpus with pre-defined DAs. We further propose to filter out the examples that might be incorrect by two confidence measures. Experimental results show that the proposed method is an efficient approach to create new dialog act corpora at low costs.

The following section presents a short review of semi-automatic labeling approaches. Section 3 describes the first stage of corpus preparation, while section 4 presents the semi-automatic training algorithm. Section 5 evaluates the approach on a French broadcast news database. In the last section, we discuss the results and we propose some future research directions.

2. SHORT REVIEW OF SEMI-AUTOMATIC DIALOG ACT LABELING APPROACHES

Semi-supervised training procedures aim at improving classifiers by retraining them on large and unlabeled corpora, as shown in [3]. The most common method to achieve this is based on the Expectation-Maximization algorithm for maximum likelihood estimation problems with incomplete data [4]. Basically, this process iteratively infers the unknown labels and retrains the classifier with these new labels.

EM for automatic DA labeling is used for example in [5]. Venkataraman further shows in [6] that prosodic features (duration, energy and pitch) can reduce the tagging error rate in the labeling process. Prosodic features, especially the duration of pauses between consecutive words, is used in [7] to improve the performance of speech segmentation into DAs.

An unsupervised approach is also shown in [8], where a set of superficial features is used to classify utterances into DAs. This set contains features that can be extracted automatically from the corpus, such as the presence or absence of *wh-words* and *question marks*. Every DA is then characterized from these features: for instance, *why questions* are defined by a *wh-word* at the initial position of an utterance, and *yes/no questions* contain a finite verb at the first position and a subject at the second position. Unsupervised classification is carried out by maximizing the *posterior* probability. With this method, no manual labeling is required, but it may be difficult to be used with any kinds of application and DAs sets.

3. CORPUS PREPARATION

This work makes use of part of the ESTER corpus [9], which contains transcribed French broadcast news speech. The chosen set of dialog acts is based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [10], where 42 dialog acts classes are defined. This list is usually reduced for recognition into a much smaller number of broad classes, depending on the application and available corpus.

Our initial DAs set contained 24 DAs that occur in the ESTER corpus, including a "radio specific DA", which represents statements specific to the radio application, such as: "France Inter, il est 5 heures" (*France Inter, it's 5 o'clock*). This set is further reduced down to 7 classes, by grouping together some classes that have very few observations (e.g. accepts and agreements).

Two subsets, composed of 12 radio emissions each, are first selected randomly and manually segmented and labeled with DAs: the first set is the initial manual training set, and the second set is used for testing. Their composition is shown in table 1.

No.	Clustered DA Type	Tag	Training		Test.
			Man.	Init.	Man.
1.	Statements	s	251	251	609
2.	Yes/No question	qy	24	24	27
3.	Other questions	q	39	527	72
4.	Dialog delimitations	oc	55	446	23
5.	Agreements	a	44	65	34
6.	Backchannel-hesitations	h	46	148	71
7.	Radio specific DAs	g	130	191	93
Tot.	All DAs		589	1652	929

Table 1. Structure of initial corpus for semi-automatic labeling.

The 1652 DAs in the initial training corpus consists of 589 DAs labeled manually plus 1063 DAs labeled automatically using rules. These rules are defined manually, based on general properties of the French language. Examples of rules are: every utterance starting with "est-ce que" is an *yes/no question*; or every utterance starting with a wh-word (such as "comment", "combien", ..) is a *why-question*, etc. The unlabeled part of corpus is composed of 5230 utterances.

4. SEMI-AUTOMATIC CORPUS LABELING APPROACHES

We describe next respectively our dialog act models, the semi-supervised training algorithm and the proposed confidence measures used to filter out the ambiguous training examples.

4.1. Dialog acts modeling

Each dialog act is represented by a unique state in the ergodic HMM shown in figure 1. Each state computes the observation log-likelihood from a unigram model described in equation 1.

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

where C encodes the dialog act class and w_i represents the i^{th} word of the current utterance.

Transition between states encode transition probabilities between subsequent dialog acts. In the following experiments, these transitions are set equiprobably between every DA-pair, with a loop probability that models the average duration of all DAs on the training corpus.

Unlike our previous works in automatic DA recognition, prosodic information is not included in the feature vector: DA models exploit lexical features only. This choice has been made to simplify as much as possible the models and parametrization domain to be used in the EM procedure. Furthermore, because of the small size of the initial corpus, only unigram statistics are computed. Obviously, once a larger part of the corpus has been semi-automatically labeled, this simplified framework can be advantageously replaced by more complex models, with prosodic features and longer temporal dependencies for example. But the most critical part of the corpus creation process is certainly just after initialization, which is the focus of this work.

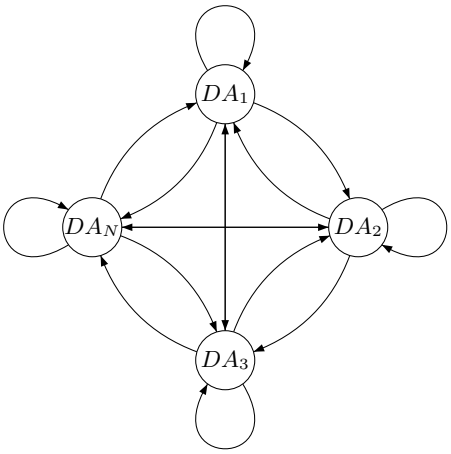

Fig. 1. Dialog act model: each node of the ergodic HMM represents one DA class.

4.2. Semi-supervised training

A small initial corpus is manually segmented and labeled with the dialog acts listed in table 1. The rest of the corpus is not initially segmented nor annotated with dialog acts. On this part of the corpus, we consider that the labels (the DA classes) are instances of an hidden random variable C. This variable is estimated by the classical Expectation-Maximization algorithm, as follows:

1. Initialization: let $\mathcal{D}_0$ be the small initial training corpus manually labeled, and Ω be the complete corpus (labeled or not); let $t = 0$;
2. The classifier is trained on $\mathcal{D}_t$;
3. The DAs of the unlabeled corpus $\Omega - \mathcal{D}_t$ are infered (and segmented) by the current classifier;
4. For each recognized DA, a confidence measure is computed to assess its reliability; let $\mathcal{M}_t$ be the most reliable DAs;
5. The most reliable examples are included into the training corpus: $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{M}_t$;
6. t is incremented, and the procedure is iterated from step 2 until a given number of iterations is reached.

4.3. Dialog act recognition

The performance of the classifier is evaluated at each iteration on the test corpus, which has been manually segmented and labeled. Recognition is realized with the ergodic HMM of figure 1 and the Viterbi algorithm, which outputs both the DA labels and their temporal limits. Recognition rate is computed for each word by comparing the recognized and correct labels.

4.4. Confidence measure

Like many confidence measures used in speech recognition [11], our first confidence measure for DA recognition is an estimate of the *a posteriori* class probability. The output of our lexical classifier is $P(W|C)$, where C is the dialog act class and W is the words sequence in the DA. The likelihoods $P(W|C)$ are normalized to compute the *a posteriori* class probabilities:

$$P(C|W) = \frac{P(W|C).P(C)}{\sum_{D \in \mathcal{DA}} P(W|D).P(D)} \quad (2)$$

$\mathcal{DA}$ is the set of all DAs and $P(C)$ is the *prior* probability of the DA class C.

In the first version of our training algorithm, called *maximum a posteriori probability* method, only the DAs $\hat{C}$ so that

$$\hat{C} = \arg\max_{C}(P(C|W)) \quad (3)$$

$$P(\hat{C}|W) > T \quad (4)$$

are included into the training corpus.

In the second version, called *a posteriori probability difference* method, the difference between the *best* hypothesis and the *second best* one is computed by the following equation:

$$P\Delta = P(\hat{C}|W) - \max_{C \neq \hat{C}}(P(C|W)) \quad (5)$$

Only the DAs with $P\Delta > T$ are included into the training corpus. This second approach aims at identifying the DAs that "dominate" all the other candidates, which is not always well captured by the first measure.

T is in the both cases an acceptation threshold and its optimal value is found experimentally.

5. EXPERIMENTS

In the following experiments, the unigram probabilities $P(w_i|C)$ with less than 6 examples in the training corpus are smoothed to the class-independent backoff *prior* $P(w_i)$. Furthermore, all DA *priors* are set equiprobable, because the training corpus is generated partly from hand-crafted rules that bias the estimates of these *priors*.

5.1. Maximum *a posteriori* probability

Figure 2 plots the DA recognition rate on the manually labeled test corpus, with the maximum *a posteriori* probability method, in function of the number of EM iterations and for different values of T. The results obtained without any confidence measure (or equivalently for $T = 0$) are also shown with the label "EM". We can note that the performance of this EM-only curve degrades, which justifies the use of confidence measures to filter out incorrectly recognized DAs.

After three iterations, the recognition rate tends to stabilize, with a maximum at 80 % for threshold 0.999 and at the third iteration. The improvement due to our semi-supervised training algorithm represents a decrease of 30 % of the recognition errors. The evolution of the size of the training corpus is shown in figure 3.

Table 2 shows the recognition rate per DA at different iterations with $T = 0.999$. One can observe that most of DA rates increase. Only the score of "yes/no questions" is decreasing , which is probably due to the lack of training data for this class in the initial manual corpus.

5.2. *A posteriori* probability difference

The DA recognition rate in function of the number of EM iterations is shown in figure 4. The corresponding corpus sizes are shown in figure 5.

The results stabilize after the seventh iteration, with a maximum of 78 % for threshold 0.9995: this represents a decrease of 27 % of the recognition errors.

Because of the very high absolute values of the thresholds retained, the difference between the Maximum *a posteriori* probability and the *A posteriori* probability difference methods is not very important.

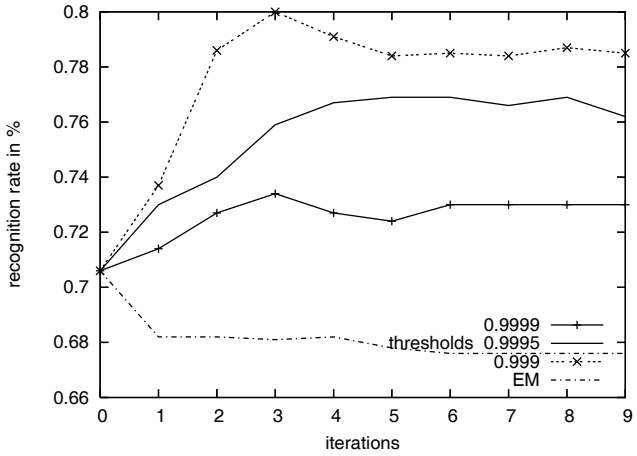

Fig. 2. Performance of the maximum *a posteriori* probability method: the X-axis represents the number of EM iterations and the Y-axis plots the DA recognition rate.

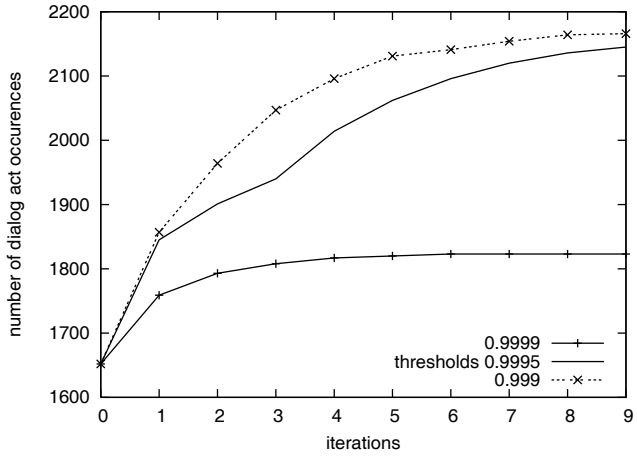

Fig. 3. Performance of the maximum *a posteriori* probability method: the X-axis represents the number of EM iterations and the Y-axis plots the DA corpus size.

6. CONCLUSIONS

The main contribution of this work is to instantiate the general EM procedure to the task of creating new semi-supervised corpora labeled with different sets of dialog acts and in different languages at a low cost. We show that confidence measures are required to filter out incorrect examples, and we evaluate two such measures on this task. Furthermore, we describe how our dialog act recognition system, which was previously developed for a Czech reservation application, can be retrained and successfully adapted to a new language (French), a new type of corpus (broadcast news) and a different set of dialog acts.

The perspectives of this work are numerous, including the evaluation of the method on corpora that are not transcribed in words (which requires to pre-process the signal with an automatic speech transcription system), the use of more complex dialog act models (for instance with prosody and dialog grammars), the development of better confidence measures and initial dialog act rules and the use

	Recognition rate in [%]							
Iter.	s	qy	q	oc	a	h	g	glob.
0	72.4	70.3	62.9	66.1	51.4	100	41.6	**70.6**
1	76.4	58.6	62.9	66.1	51.4	100	42.7	**73.7**
2	81.8	58.0	62.5	66.1	65.3	100	45.7	**78.6**
3	83.8	52.3	65.5	66.1	65.3	100	41.0	**80.0**
4	82.6	51.1	66.5	66.1	62.5	100	43.1	**79.1**
5	81.9	47.1	68.2	66.1	62.5	100	43.1	**78.4**
6	81.8	51.1	68.2	66.1	62.5	100	43.1	**78.5**
7	81.8	46.8	68.2	66.1	62.5	100	43.1	**78.4**
8	82.2	46.8	68.8	66.1	62.5	100	43.1	**78.7**
9	81.9	46.8	68.8	66.1	62.5	100	43.1	**78.5**

Table 2. Performance of the maximum *a posteriori* probability method: dialog acts recognition rate in % at different iterations with probability threshold 0.999.

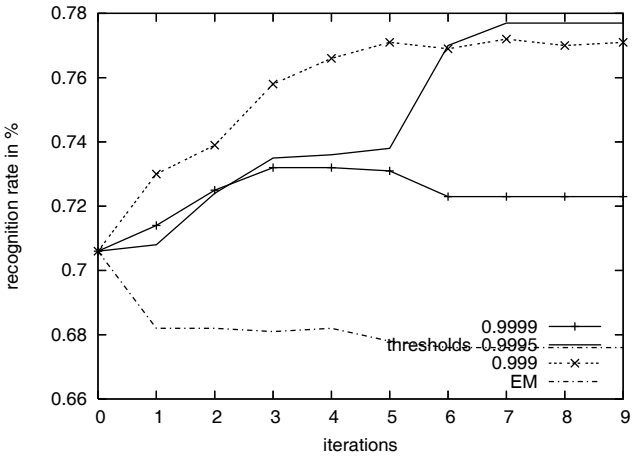

Fig. 4. Performance of the *a posteriori* probability difference method: The X-axis represents the number of EM iterations and the Y-axis plots the DA recognition rate.

of more advanced filtering strategies such as in active learning.

7. ACKNOWLEDGEMENTS

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).

8. REFERENCES

[1] 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.

[2] P. Král, J. Klečková, and C. Cerisara, "Automatic Dialog Acts Recognition based on Words Clusters," in *WESPAC IX 2006*, Seoul, Korea, June 2006.

[3] K. Nigam *et al.*, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, pp. 1–34, 1999.

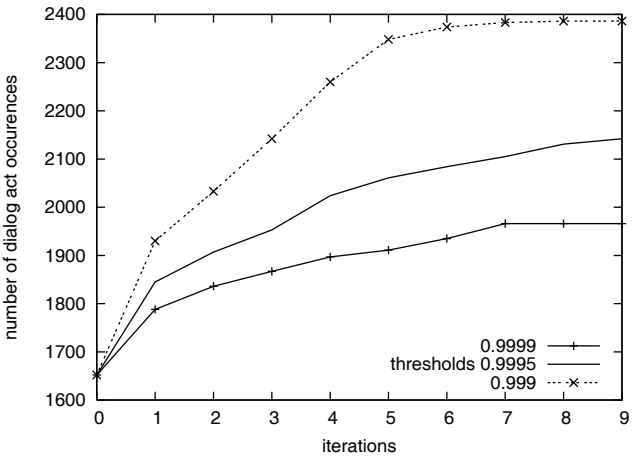

Fig. 5. Performance of the maximum *a posteriori* probability method: The X-axis represents the number of EM iterations and the Y-axis plots the DA corpus size.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. 1, no. 39, pp. 1–38, 1977.

[5] A. Venkataraman, A. Stolcke, and Shriberg E., "Automatic Dialog Act Labeling with Minimal Supervision," in *Australian International Conference on Speech Science and Technology*, Melbourne, Australia, December 2002, Australian Speech Science and Technology Association.

[6] A. Venkataraman, L. Ferrer, A. Stolcke, and Shriberg E., "Training a Prosody-based Dialog Act Tagger from Unlabeled Data," in *ICASSP'03*, Hong Kong, April 2003, vol. 1, pp. 272–275.

[7] M. Zimmermann, A. Stolcke, and E. Shriberg, "Joint segmentation and Classification of Dialog Acts in Multiparty Meetings," in *ICASSP'06*, Toulouse, France, May 2006, pp. 581–584.

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

[9] Département Technologies de l'Information et de la Communication Action Technolangue, "French ESTER Corpus," in *http://www.recherche.gouv.fr/technolangue/*.

[10] 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.

[11] E. Lleida and R. C. Rose, "Likelihood Ratio Decoding and Confidence Measures for Continuous Speech Recognition," in *ICSLP'96*, Philadelphia, USA, 1996, vol. 1, pp. 478–481.