論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Speaker Adaptation for Dialog Act Recognition
著者(和文)	ﾖﾊﾝ ﾛﾀﾞﾝ, 篠田 浩一
Authors(English)	Johan Rohdin, Koichi Shinoda
出典(和文)	日本音響学会2012年春季研究発表会講演論文集, Vol. , No. , p. 111
Citation(English)	2012 Spring Meeting ASJ, Vol. , No. , p. 111
発行日 / Pub. date	2012, 3

Speaker Adaptation for Dialog Act Recognition *

◎ Johan Rohdin and Koichi Shinoda (Tokyo Institute of Technology)

1 Introduction

A dialog act (DA) describes the purpose or role of an utterance and is important for language understanding. Typical examples of DA classes are *Statement* or *Backchannel*. Applications of DA recognition systems are meeting summarization [1] and constraining speech recognition hypothesis [2].

It could be expected that there are some speaker specific patterns in the features that characterize different dialog acts, in which case a DA system may benefit from speaker adaptation.

Speaker adaptation proved to be effective for improving DA *segmentation* in [4]. In this study we will investigate the effect of speaker adaptation for *joint* segmentation and classification, i.e., DA recognition. We propose using maximum a posteriori (MAP) adaptation for DA systems based on conditional random fields (CRF). CRFs have so far performed the best for joint DA recognition [3]. MAP adaptation for maximum entropy models (MEMs) was proposed in [5]. The extension of that method to CRFs is straightforward.

2 Data and Labeling scheme

We used the ICSI (MRDA) [8] meeting corpus which consists of naturally occurring meetings. There are 51 meetings in the training set, 11 in the test set, and 11 in the development set. The training set contains 530k words in 82k dialog act tagged segments. For the evaluated speakers, the number of words in the adaptation data varied from 236 to 106k. We used the classes: *Statement* S, *Question* Q, *Backchannel* B, *Disruption* D, *Floor mechanism* F, and *Unclassified* Z.

In [3], five labeling schemes for using CRF for joint DA recognition was proposed. These schemes label every word. We used the coding scheme denoted EI in this paper. The EI scheme uses two labels for each DA class: one for the final word of a DA and another for any other words in a DA segment.

3 Conditional Random Fields

3.1 Model description

A linear-chain conditional random field (CRF) [6] estimates the conditional probability of a label sequence $\boldsymbol{y} = y_1, \ldots, y_T$ given an observation sequence $\boldsymbol{o} = o_1, \ldots, o_T$ by

$$P_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{o}) = \frac{1}{Z_{\boldsymbol{o}}} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k\left(y_{t-1}, y_t, \boldsymbol{o}, t\right)\right), \tag{1}$$

where index k indicates the k-th feature. The weights λ_k's are typically estimated by maximizing the conditional likelihood, $P_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{o})$.

3.2 Maximum a posteriori adaptation for CRF

In MAP adaptation of each feature weight in CRF, we used a Gaussian prior whose mean is equal to the corresponding weight of the speaker-independent CRF. This choice of priors was first proposed in [5] for MEMs. This gives the log posterior

$$l\left(\boldsymbol{\lambda}; \boldsymbol{o}\right) = \sum_{j=1}^{J} \log\left(P_{\boldsymbol{\lambda}}\left(\boldsymbol{y}^{(j)}|\boldsymbol{o}^{(j)}\right)\right) - \sum_{k=1}^{K} \frac{\left(\lambda_k - \lambda_k^0\right)^2}{2\sigma^2}, \tag{2}$$

where λ_k^0 is the weight for feature f_k of the original model, λ_k is the weight of feature f_k of the adapted model to be estimated, and σ is the standard deviation of the prior. The index j indicates each training instance. In this study we use the L-BFGS algorithm which needs a closed form expression of the gradient of the log posterior. As can be seen in Eq. (2), changing the mean of the prior changes the gradient in a trivial way.

4 Experiments

4.1 Experimental conditions

Every word sequence to be labeled corresponds to one person's speech from one whole meeting. We only considered reference conditions, i.e., using the words from the transcripts of the corpus.

We trained four different kinds of models. The first model was trained using all data in the training

* 発話行為認識のための話者適応　ロディーンヨハン、　篠田浩一　（東工大）

Table 1 Individual DA results and their total frequency in the test and development set.

DA class		Z	S	Q	B	F	D	Overall
Number of instances		414	17915	2222	3958	3705	4572	32786
F-measure (%)	SI	7.2	34.5	20.2	68.0	35.5	12.7	35.3
	ALL	8.2	35.2	22.2	68.7	37.0	13.3	36.1
	MAP_SI	11.6	35.8	24.7	69.3	38.4	14.3	36.9
	MAP_ALL	7.9	35.7	23.3	69.2	38.6	13.7	36.7
Strict (%)	SI	97.2	71.3	91.2	36.0	72.3	94.5	75.1
	ALL	97.2	70.8	89.9	35.2	70.6	94.2	74.5
	MAP_SI	95.2	70.1	88.1	34.6	69.7	93.1	73.7
	MAP_ALL	97.5	70.3	89.5	34.8	69.4	94.3	74.1

set. This model is to some extent speaker-dependent since all the speakers in the test and development set are involved in the training set. We call this model *ALL* model. We also trained a speaker-independent (SI) model for each speaker by using all data in the training set except the data from the same speaker.

Finally we used MAP adaptation from both ALL and SI models for each speaker in the test and development set, using the speakers' data in the training set as adaptation data. We refer to these models as MAP_ALL and MAP_SI respectively.

Except that we did not use any prosodic features, we used the same features as in [3], namely, word unigrams, bigrams, and trigrams, in the context of ± 2 words and label bigram features.

We used both the development set and the test set for the evaluation and chose the variances of the priors used for training ALL and MAP_ALL models respectively, by cross validation. For SI and MAP_SI models, we used the same variances as for ALL and MAP_ALL respectively.

We used the Wapiti toolkit [9] and modified the gradient calculation for MAP adaptation.

4.2 Evaluation metrics

We used the F-measure [3] and the *Strict* metric [10]. These metrics consider a DA segment to be correctly recognized only if both segmentation and classification are correct. The Strict metric then measures the percentage of the words that occur in an incorrectly classified DA segment. In the calculation of F-measure, every DA segment was counted as one unit.

4.3 Results

The results are shown in Table 1. The ALL model outperformed the SI models. MAP adaptation from the ALL model improved the performance further. Comparing the MAP_ALL and the SI models, we observed a relative gain in F-measure of 4.1%.MAP adaptation from the SI models seemed to give slightly better result than from the ALL model. The results varied significantly among the different DA classes.

5 Conclusion and future work

In this paper we have proposed a MAP adaptation method to speakers for dialog act systems based on conditional random fields. We evaluated the method on the ICSI meeting corpus and found that MAP speaker adaptation gives significant improvements.

Future work should include more sophisticated adaptation methods, integration with speech recognition as well as using more kinds of features. Particularly, prosodic features would be interesting since this was not considered in this study.

References

[1] Murray *et al.*, Proc. HLT-NAACL, 367-374, 2006

[2] Ji and Bilmes, in Proc. ICASSP, 5110-5113, 2010

[3] Zimmermann, INTERSPEECH, 864-867, 2009

[4] Kolář *et al.*, Speech Communication, 52 (3), 236 - 245, 2010,

[5] Chelba *et al.*, Computer Speech & Language, 20 (4), 382 - 399, 2006

[6] Lafferty *et al.*, In Proc. ICML, 282–289, 2001

[7] Chen *et al.*, IEEE Trans. on Speech and Audio Processing, 8 (1), 37 -50, 2000

[8] Shriberg *et al.*, In Proc. 5th SIGdial Workshop on Discourse and Dialogue, 97–100, 2004

[9] Lavergne *et al.*, Proc. ACL, 504–513, 2010

[10] Ang *et al.*, Proc. ICASSP '05, 1061 - 1064, 2005