Vol.2011-SLP-87 No.8
2011/7/22

Speaker Adaptation for Dialogue Act Classification

Johan Rohdin[†1] and Koichi Shinoda[†1]

In this paper we investigate MAP adaptation to speakers for dialog act classification systems based on conditional random fields. MAP adaptation is done by assuming a Gaussian prior of the model-weights with mean equal to the weights of a baseline model. We did experiments on the ICSI meeting corpus and found that speaker adaptation gives significant improvements of the dialog act classification accuracy.

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 automatic DA recognition systems are meeting summarization[1] as well as constraining speech recognition hypothesis of spontaneous speech[2]. Such applications require both segmentation and classification of a word-stream into dialog acts. Recent studies[3][4] suggests that doing segmentation and classification jointly (i.e., DA recognition) is preferably to doing it sequentially. If the word transcrpition is not known, the DA segmentation and classifiaction process should ideally be integrated with the speech recognition system. In this study, we will use reference transcripts, i.e., not the output of a speech recognizer, as proper integration of DA recognition and speech recognition is a difficult topic in its own that yet remains to be solved.

It could be expected that there are some speaker specific patterns in the features that characterize different dialogue acts. For example, a specific speaker may often start his/her questions with the phrase *I wonder...* and another with *can I ask...*. As another example, different speakers may prefer to use different phrases in order to take the *floor* (floor-grabbers), such as *Um*, *so* or *but*. Due to these speaker specific patterns, a DA system is expected to benefit from speaker adaptation.

Kolář et al.[5] showed that speaker adaptation is useful for the *segmentation* task. A natural question is whether it is also useful for *classification* and/or *joint* segmentation and classification. This have not yet been investigated. The study by Kolář et al.[5] considered speaker adaptation of two different systems for the *segmentation* task. One system was based on decision trees with various prosodic features and the other was based on a hidden event language model (i.e., a system that uses only word features). They found small but statistically significant improvements by speaker adaptation for both systems, both when using reference transcripts and when using the 1-best ASR hypotheses. The improvements were larger for the system that used word features than for the system that used prosodic features.

In this study we will focus on speaker adaptation in DA classification. Even though most applications may require both segmentation and classification, we believe it is also important to investigate to what extent the segmentation and classification tasks benefits from speaker adapatation individually, in order to formulate a suitable adaptation scheme for the joint task.

We will use CRF in these experiments and propose using Maximum a posteriori (MAP) adaptation[7] for custumizing the model to speakers. As far as we know, Conditional Random Fields (CRF) have so far performed best for joint DA segmentation and classifiaction[4]. We will use word features as well as a DA *boundary indicator* variable but we will not use any prosodic features.

We are not aware of any studies that consider adaptation methods for DA sytems based on CRF, but there are some studies of domain and/or speaker adaptation of related tasks and models. For example, MAP adaptation for maximum entropy models (MEMs) was proposed by Chelba et al.[7] for domain adaptation of text capitalization. The extension of that method to CRF is straightforward. Quite few other methods have been proposed for adaptation of MEMs or CRF such as the the mega model[8], a feature augmentation scheme[9] and hierarchical Bayesian domain adaptation[10]. These methods were applied for domain adaptation of various language processing tasks e.g. named entity recognition or capitalization. MAP as well as a maximum conditional likelihood linear regres-

[†1] Tokyo Insitute of Technology

sion (MCLLR) technique for speaker adaptation of acoustic models for hidden conditional random fields (HCRF) has alos been investiagted[11].

Section 2 of this paper describes the task of DA classification. In Section 3, the CRF framework is explained. In Section 4, we explain the experimental conditions and results. In Section 5 we discuss the method and the results. Finally, our conclusion and some ideas for future work are given in Section 6.

2. Problem description

2.1 DA classification

In the DA classification task, a sequence of words and a segmentation of these words into DA segments are given. The task is then to asign a DA label to each of these segments. This task is simplified from the one of most applications where neither the transcribed seqeuence of words nor its segmentation into DA segments are available. In that situation we have to rely on a speech recognizer to obtain the words, and the segmentation into DA segments must also be found automatically.

A large variety of features could be considered for the DA classification task. For example, word N-grams either existing at a specific position or anywhere in the segment and/or the number of words in the unit. Also prosodic features (pitch, energy, duration and pauses) are useful[14].

Typically, DA sytems also considers transition probabilities between different DAs. For example, the probability that Question is followed by Statement. For that purpose, we have the choice to sort the DA segments either by time regardless of who is the speaker, or to treat every speaker separatly. We will use the second approach in this study (see discussion in Section 5).

2.2 Data

We used the ICSI (MRDA)[15] meeting corpus which consists of naturally occuring meetings, 51 in a training set 11, in a test set and 11 in a development set. In addition to word transcripts, the corpus is annotated with a detailed set of DA classes. In order to reduce the number of classes, several different *classmaps* are provided. We used the classmap called 01b in the corpus which has six classes: *Statement* S, *Question* Q, *Backchannel* B, *Disruption* D, *Floor mechanism* F, and *Unclassified* Z. The corpus also provides the opportunity to choose whether to split segments that are prosodically one segment but syntactically two. Since we focus on word features, we choose to split such segments. In our experiments we use speaker specific data in the training set to do adaptation for all speakers in the test set and the development set (see Section 4.1). Two speakers in the development set were excluded from the experiments since they had no data in the training set. With this set-up, the training set contains 530k words in 82k dialog act tagged segments. For the speakers for which adaptation was done, the number of words in the adaptation data varied from 236 to 106k and the number dialog act tagged segments from 48 to 14k. Table 1 shows an example from the corpus.

Table 1 An example of word stream from one speaker, the *observed* border variable (see Section 4.1), and the correponding DA label.

Words Boundary	*completely* False	*irrelevant* True	*yeah* True	*for* False	*nois-* False	*noise* False	*cancelling* True	*Um* True
DA label	S	S>	B>	S	S	S	S>	F>

2.3 Labeling scheme for CRF

Zimmermann[4] suggested five coding schemes for using CRF for joint DA segmentation and classification. These coding schemes labels every word instead attaching one label a whole DA segment. We will use the coding scheme denoted EI in the paper by Zimmermann. The coding scheme uses two labels for each DA class: one for the final word of a DA and one for any other words in a DA segment. For example, label S> corresponds to the final word of Statement and label S correspond to any words except the final, of Statement. This coding scheme is a good trade-off between performance and complexity.

Since this approach labels every word instead of the DA units, we cannot use features such as the number of words in the DA unit. We use this approach since it easily extends to the joint task and since we are here mostly interested in seeing the effect of speaker adaptation rather than finding the optimal set of features.

3. Conditional Random Fields

3.1 Model description

A linear chain Conditional Random Field (CRF)[12] estimates the conditional probability of a label sequence $\boldsymbol{y} = y_1, \ldots, y_t, \ldots, y_T$ given the observation (sequence) $\boldsymbol{o} = o_1, \ldots, o_t, \ldots, o_T$ by

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

where the index k indicates a feature. The weights λ_k are typically estimated by maximizing the (conditional) likelihood. In many language processing systems as well as in this study, the feature functions, f_k, are binary.

To avoid over-training, a suitable *prior* probability distribution for the weights can be assumed. Often a Gaussian distribution with mean zero is used[13]. Instead of a (conditional) likelihood for the weights, we then get a (conditional) *posterior* probability distribution for the weights to maximize. Its logarithm is given by

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

where the index j indicates the training instances. From Eq. (2) it can be seen that the Gaussian prior can also be interpreted as L_2-regularization of the log likelihood function.

The hyper-parameter σ is usually estimated by cross-validation. The weights that maximizes Eq. (2) cannot in general be found analytically but there are several numerical methods that can precisely estimate the parameters.

3.2 Maximum a posteriori adaptation for CRF

MAP adaptation can be done by using a Gaussian prior with the mean vector equal to the original (ML-estimated) weight vector instead of zero. This was first proposed in[7] for MEMs. Since the only difference between a linear chain CRF and a MEM is that the feature functions, f_k, of a CRF includes y_{t-1}, the same MAP adaptation method can be applied to CRF. This gives the log posterior

$$l\left(\lambda; \boldsymbol{o}\right) = \sum_{j=1}^{J} \log \left(P_{\boldsymbol{\lambda}} \left(\boldsymbol{y}^{(j)} | \boldsymbol{o}^{(j)} \right) \right) - \sum_{k=1}^{K} \frac{\left(\lambda_k - \lambda_k^*\right)^2}{2\sigma^2}, \qquad (3)$$

where λ_k^* is the weight for feature f_k of the original models and λ_k the weight of the k-th feature of the adapted model to be estimated. Notice that, according to Eq. (3), those weights not included in the adaptation data may also change by MAP adaptation. 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. (3), changing the mean of the prior changes the gradient in a trivial way.

4. Experiments

4.1 Experimental conditions

In this study we only considered reference conditions, i.e. using the words from the transcripts of the corpus and not the result of automatic speech recognition. We trained three different kinds of models. The first model was trained using all data in the training set. This model is to some extent speaker dependent since all the speakers in the test set and development set has some data in the training set. It may therefore serve as an trivial baseline for adaptation. We call this model the *All_train* model.

We also trained a speaker independent model for each speaker by using all data in the training set except the data from the specific speaker. Excluding the training data from all speakers in the test set and development set in order to make one speaker independent model would have reduced the training set too much. We refer to these models as SI models.

Finally we used MAP adaptation from the All_train model for each speaker in the development set and test set, using the speakers data in the training set as adaptation data. Since the adaptation data was already included in the training data of the All_train model, no new features will be introduced by the adaptation but the feature weights may change. We refer to these models as MAP models.

We used the same word features as Zimmermann[4], namely word unigrams, bigrams and trigrams in the context of ±2 words. We did not use any prosodic features. Instead of pure label features as in that study, we use label bigram features in combination with a *boundary indicator* observation in order to add the

Table 2 Optimal regularization parameters.

	Testset	Devset
All_train	$\sigma_t^{2*} = 1/2$	$\sigma_d^{2*} = 1/2$
Adaptation	$\sigma_t^2 = 1/60$	$\sigma_d^2 = 1/195$

segmentation information. The boundary indicator indicates whether a word is the final word of a DA segment or not. For example, it indicates *Boundary=True* at S> and *Boundary=False* at S in the training set (see Table 1). This variable is observed also in the testing phase. Ideally, CRF should learn to follow this variable and never predict S> at *Boundary=False* in the testing phase. The word sequences to be labeled corresponds to one persons speech from one whole meeting. For the first word in the sequence there is no label bigram and therefore the *boundary indicator* will not be taken into account. Therefore the above method may not always predict boundaries correctly for those words. Such cases were treated as errors in the evaluation.

We used the toolkit Wapiti[16] which we modified little in order to do MAP adaptation. Wapiti supports many methods to stop the training. We used the default value of maximum number of line-searches as the only stopping criteria.

4.2 Evaluation metric

As evaluation metric, we use the *DA error rate* (DER)[19]. This metric is intended for the joint task and considers a DA segment to be correctly recognized only if the surrounding borders are correctly identified, no additional borders are inserted in the correct segment, and the segment is given the correct label. Since the segmentation is given in the classification task, the first criteria should not be violated. As mentioned in the Subsection 2.3, the framework we use may fail to correctly input DA borders in rare occasions.

4.3 Cross validation

We used both the test set and the development set for the evaluation. The parameters for the priors were optimized by 2-fold cross validation in the following way:

(1) Find optimal hyper-parameter for the All_train model, σ_t^* and σ_d^*, for the test set and development set respectively.

(2) Use the weights, λ_k^*, from the model trained with σ_t^* for MAP adaptation (see Eq. (3)) and find the optimal parameter σ_t for the test set. Likewise, use the weights from the model trained with σ_d^* to find σ_d.

(3) Finally, evaluate the test set using the parameter pair (σ_d^*, σ_d) and evaluate the development set using (σ_t^*, σ_t)

Table 3 Oversall speaker adaptation results.

	SI	All_train	MAP
DER (%)	22.6	22.2	22.0

The optimal values for σ^2 are given in Table 2. We optimized $1/\sigma^2$, which is the parameter to specify in Wapiti (ρ_2), in steps of 1 from 1 to 8 for the All_train models and using the values $5, 10 \ldots, 160, 170, \ldots, 200, 220, \ldots, 400, 450. \ldots, 1000$ for MAP adaptation. If two of the values gave the same result we used their average. For training the SI models, we used the same parameter value as for the All_train models.

4.4 Results

The results are shown in Table 3.

The DER decreased from 22.6% for the SI models to 22.2% for the All_train model that was trained with all speakers' training data. MAP adaptation decreased DER further to 22.0%. The improvement of MAP adaptation from the All_train model is statistically significant for $p = 0.026$ by a two tailed Sign test that compares the two systems' predictions of each DA segment.

For further analysis the results for each DA class are shown in Table 4. The *unlabeled*, Z, is included here although one could question whether the system should be allowed to classify a segment as unlabeled.

The overall improvement by MAP adaptation is quite modest. However, looking at the individual DA results, we can see that all DA classes except statements and the unlabeled class has much larger improvements. Statements showed no improvement from speaker specific modelling (i.e., All_train or MAP) compared to using speaker independent models. Since more than half of the instances are statements, the overall improvement is low.

5. Discussion

Since Backchannels and Floorgrabbers have no syntactic meaning, there might

Table 4 Individual DA results and their frequency in the test and development set.

DA class		Z	S	Q	B	F	D
Frequencey		414	17915	2222	3958	3705	4572
DER (%)	S.I.	75.1	8.2	53.4	30.2	26.1	50.0
	All_Train	76.3	8.2	50.0	29.9	25.4	49.2
	MAP	77.5	8.2	49.5	29.7	24.8	48.8

be more room for every individual speaker to choose their own voacbulary compared to statements where the voacbulary must be chosen so that it transfer the correct message. However, also Questions have some constraints on the choice of vocabulary but still seems to benefit siginficantly from speaker specific modelling. The improvements of Disruptions might be because some speaker tend to have more interupted utterances overall or that a certain kind of utterances are often interupted for a specific speaker. However, much more analysis is needed before any conclusions like this can be made. The poor ressults for statements might also be explained by the fact that the error rate for them is very low compared to other classes and therefore might be difficult to improve.

For applications it would obviously be desired to increase the improvements of speaker adaptation, especially considering the efforts needed in order to annotate adaptation data.

As mentioned in Section 4, the amount of adaptation data varied greatly among the speakers. It seems natural that adaptation would work better the more adaptation data available. This issue has not been investigated in this study. Just comparing the improvements for the individual speakers in this study may not be sufficient because any differences among them may not only depend on the amount of adaptation data but also on how their speaking style fits the All_train model.

It should also be remembered that the training data for the All_train models in this study included the adaptation data and therefore cannot be said to be speaker independent. If the amount of adaptation data become very large, such All_train and MAP adapted models would become more similar. Therefore, when investigating how much adaptation data is needed, we should compare with the speaker independent models even though the amount of training data for them varied among the speakers in this study. We also tried do MAP adaptation from the speaker independent models which is how MAP adapation is typically done but we did not obatain any good results in this way. One explanation might be that we used the regularization that was found optimal when doing MAP adaptation from the All_train models.

As mentioned in Section 2.3 we have the choice to sort the DAs either by time regardless of who is the speaker or to treat every speaker separately. The first approach requires that segmentation into DAs or at least into speaker turns are given as is the case for the classification task. Segmenting into speaker turns is trivial if there is no ovelapping speech but this is rarely the case in reality. For example, one speaker could utter a backchannel in the middle of another speakers statement. Also, for adaptation experiments the first approach will be a bit more tricky also for the classification task since it does not allow us to use a specific CRF for every speaker. The problem could be overcome by using speaker ID as feature. In this study we chose the second approch since, as mentioned, we want the extension to the joint task to be easy.

6. Conclusion and future work

In this paper we have proposed a MAP adaptation method to speakers for dialog act classification systems based on CRF. We evaluated the method on the ICSI meeting corpus and compared to speaker independent models and baseline model made by including the adaptation data in the training set. We found that MAP adaptation gives statistically significant improvements from such baseline model. The overall DA error rate (DER) decreased from 22.6% for the speaker independent models to 22.2% for the speaker dependent baseline model. MAP adaptation decreased the DER further to 22.0%. Larger improvements were observed for four out of six individual DA classes.

The improvements were statistically significant but not so large. It would be highly desirable that future work lead to larger improvements. There are several possibilities that can be investigated. As mentioned in Section 5, the amount of adaption data needed is not yet clear. Use of proper amount of adaptation data may improve the results in this study.

Some improvements of the adaptation method might also be possible. We used one value for the hyper-parameter, σ^*, for training the baseline model, and one value for the hyper-parameter, σ, for MAP adaptation. Instead, we can choose different hyper-parameters for different groups of weights both when training the baseline model and for MAP adaptation. For example, one for all word unigram features and another for all word bigram features etc.. As mentioned in the introduction, there are also a few other adaptation schemes that has performed better than domain MAP adaptation of MEM for various language processing tasks[8),9)10)]. Also, as can be seen in Table 2, the optimal adaptation parameters differed significantly for the two sets and a better estimation of the parameters, e.g. by using more sets in the cross-validation, may improve the results.

In this study we used only word features. Naturally it would be interesting to investigate speaker adaptation of various prosodic features too. The integration with a speech recognizer was also not considered. As human-transcribed speech may not be available in many applications, this is an important area to investigate.

References

1) Murray, G., Renals, S., Moore, J. and Carletta, J. “Incorporating speaker and discourse features into speech summarization” In Proc. HLT-NAACL, New York, New York City, USA (2006)

2) Ji, G. and Bilmes, J. “Jointly recognizing multi-speaker conversations” In Proc. ICASSP, Dallas, USA, (2010)

3) Zimmermann, M., Liu, Y., Shriberg, E. and Stolcke, A. “A* based joint segmentation and classification of dialog acts in multiparty meetings” In proc. 9th ASRU, San Juan, Puerto Rico, 2005, pp 215-219

4) Zimmermann, M. “Joint segmentation and classification of dialog acts using conditional random fields”, In: Proc. INTERSPEECH 2009, Brighton UK.

5) Kolář, J., Liu, Y. and Shriberg E. “Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech”, Speech Communication 52 (2010) 236-245.

6) Guz, U., Tur, G., Hakkani-Tür, D. and Cuendet, S. “Cascaded model adaptation for dialog act segmentation and tagging” Computer Speech and Language 24 (2010) 289-306

7) Chelba, C. and Acero A. “Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot”. In: EMNLP (2004).

8) Daumé III, H. and Marcu, D. “Domain adaption for statistical classifiers.” Journal of artificial Intelligence research, (2006), 26:101-126.

9) Daumé III, H. “Frustratingly easy domain adaptation.” In Proceedings of ACL, pages 256-263, (2007).

10) Finkel, J. R. and Manning, C. “Hierarchical Bayesian domain adaptation.” In Proceedings of HLT-NAACL, pages 602-610, (2009).

11) Sung, Y., Boulis, C. and Jurafsky D. “Maximum conditional likelihood linear regression and maximum a posteriori for hidden conditional random fields speaker adaptation” In Proc. ICASSP, Las Vegas, USA, (2008)

12) Lafferty, J., McCallum, A., and Pereira, F. “Conditional random fields: Probabilistic models for segmenting and labeling data.” In Proceedings of the International Conference on Machine Learning (ICML), (2001).

13) Chen, S. F. and Rosenfeld, R. “A survey of smoothing techniques for me models.” IEEE Transactions on Speech and audio Processing, (2000) 8(1):37-50.

14) Shriberg, E. et al. “Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech” Language and Speech 41(3-4): 439-487. *Special Issue on Prosody and Conversation*, 1998

15) Shriberg, E. et al. “The ICSI meeting recorder dialog act (MRDA) corpus,” In: Proc. SIGDIAL, Cambridge, USA, 2004, pp. 97-100.

16) Lavergne, T., Cappé, O. and Yvon, F. “Practical Very Large Scale CRFs” Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL) 504-513, Uppsala, Sweden (2010)

17) Shriberg, E., Stolcke, A., Hakkani-Tür, D. and Tür, G. “Prosody-based automatic segmentation of speech into sentences and topics.” Speech Communication 32 (1-2), 127-154.

18) Ang, J., Liu, Y. and Shriberg, E. “Automatic dialog act segmentation and classification in multiparty meetings” In Proc. ICASSP, Philadelphia, USA, (2005)

19) Zimmermann M., Liu, Y., Shriberg, E. and Stolcke, A. “Toward Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings” In Proc. 2nd MLMI, Edinburgh, UK, 2005.