Joint Segmentation and Classification of Dialog Acts using Conditional Random Fields

Matthias Zimmermann

xbrain.ch, Switzerland

matthias.zimmermann@xbrain.ch

Abstract

This paper investigates the use of conditional random fields for joint segmentation and classification of dialog acts exploiting both word and prosodic features that are directly available from a speech recognizer. To validate the approach experiments are conducted with two different sets of dialog act types under both reference and speech to text conditions. Although the proposed framework is conceptually simpler than previous attempts at segmentation and classification of DAs it outperforms all previous systems for a task based on the ICSI (MRDA) meeting corpus.

Index Terms: speech analysis, dialog acts, segmentation and classification

1. Introduction

Spoken language technology research moves toward tasks beyond the transformation of the speech signal into a stream of words. For this, further processing of recognizer output becomes mandatory. Example domains that require a segmentation of the stream of words into meaningful typed units such as dialog acts may be found in discourse understanding [1], action item identification [2, 3], or meeting summarization [4].

The task we are investigating is how to split a stream of words into nonoverlapping segments of text and assign mutually exclusive dialog act types to these segments. While this task description suggests a sequential solution, an approach based on joint segmentation and classification most likely performs best as knowledge of the classification might also improve the segmentation. We use the term *joint segmentation and classification* for systems that do not implement this task in the form of two independent modules running in sequence but produce their final result by taking into account information from both the segmentation and the classification. Joint segmentation and classification of dialog acts has first been addressed in [5] in the context of the Verbmobil project. More recent work regarding joint segmentation and classification using the ICSI (MRDA) corpus [6] may be found in [7, 8, 9] where a sequential approach based on both word and prosody features was initially proposed in [10]. In [7] results are also provided for the AMI meeting corpus.

Conditional random fields [11] are explicitly designed to segment and label sequences which makes them an interesting candidate for joint segmentation and classification of dialog acts (DAs). Furthermore, conditional random field (CRF) models share some interesting properties with maximum entropy models that have already been successfully applied to joint segmentation and classification of DAs [9]. One desired property is the CRF's capability to use large amounts of correlated features such as word bigrams extracted from the word stream produced by the speech recognizer.

Acknowledgment: This work is partly based on prior research at the International Computer Science Institute (ICSI), USA.

Reference	|	S	S	|	Q	Q	Q	Q	|	D	D	D	|	B	|
Coding E		n	S>		n	n	n	Q>		n	n	D>		B>	
Coding B		<S	n		<Q	n	n	n		<D	n	n		<B	
Coding EI		s	S>		q	q	q	Q>		d	d	D>		B>	
Coding BI		<S	s		<Q	q	q	q		<D	d	d		<B	
Coding BIE		<S	S>		<Q	q	q	Q>		<D	d	D>		<B>	

Figure 1: Definition of dialog act codings schemes investigated. Word labels are represented by individual letters that refer to the type of the dialog act such as Q> for the last word of a question. For dialog act boundaries the symbol | is used in the reference line.

The remainder of this paper is organized as follows. In the next section the proposed methodology is described. Experiments and obtained results are analyzed in Section 3, while conclusions and possible future work are provided in Section 4.

2. Methodology

We first introduce the investigated coding schemes that map the joint task of segmentation and classification of DAs into the framework of CRFs. Then, the used features are presented followed by a definition of the applied performance metrics.

2.1. Coding Schemes for Joint Task

The CRFs sequence learning capability is exploited by the application of the various coding schemes taking into account a label specific context for the word and prosody based features. To map the joint task of segmentation and classification of dialog acts into the CRF framework coding schemes are used as presented below. For the training of the CRFs the coding schemes assign labels to individual words that incorporate information from both DA boundaries and DA types using the available DA gold standard. In recognition mode the CRFs are applied to the word sequence of the reference transcriptions or the output word sequence from the STT module. From the label sequence produced in the CRF output DA boundaries and types can then be derived in a simple post-processing step.

The first scheme investigated is called coding E here. It exactly represents the classification used in our previous work [9]. Using this scheme the last word of each dialog act is labeled using a DA type specific boundary label such as 'Q>' for the last word of a question. All other words are mapped to the general non-boundary label 'n' (see Fig. 1 for an example). The second

coding scheme, Coding B, is motivated by the following two observations: First, the beginning of an utterance is frequently uttered more clearly than the rest of it and second, the first word can be very indicative of the DA type (as in the case of $wh-$ questions). Coding B labels the first word of each dialog act with a DA specific label and maps all other words to label 'n'. These two labeling schemes are both conceptionally simple and represent a 6-way classification in the case of *Map 01* defined in the ICSI MRDA corpus distinguishing between the five DA types statements, questions, backchannels, floorgrabbers, and disruptions. However, they do not take full advantage of the CRF capability to learn label sequences: In the case of DAs that contain more than a single word, the CRF model "forgets" the DA specific label associated with the DA boundary by producing general non-boundary labels 'n'. A CRF model based on either E or B coding can therefore not effectively incorporate transition probabilities for DA types.

For coding schemes EI (and BI) the general non-boundary label 'n' is replaced by DA specific labels used for all words not corresponding to the last (first) word of a dialog act as shown in Fig. 1. This extension of coding schemes E and B comes at the cost of complexer CRF models as the number of classes rises to two times the number of DA types. Finally, for BIE coding the used labels not only indicate the type of the DA but specifically indicate the first word, the last word, and other words inside a dialog act. For this coding scheme the number of necessary classes rises to four times the number of DA types. Three classes for each DA type to differentiate the first, the last, and inner words, as well as an additional class for DAs consisting of a single word (as the single word backchannel '<B>' shown in Fig. 1). This last coding scheme is motivated by our previous finding [12] that both the first and the last words of a dialog act are essential for a good performance of text based DA classification. Alternatively to the BIE coding scheme the use of semi-Markov CRF [13] might prove an interesting option as it would be possible to combine features from both beginnings and endings of DAs. Furthermore, semi-Markov CRF would allow the use of features computed on complete DA candidates.

2.2. Features based on Text and Durational Prosody

For simplicity of implementation the approach to joint segmentation and classification of DAs investigated here is exclusively based on features that are directly produced by the available speech to text system (STT). Next to the sequence of words, discretized pause durations are exploited as in our previous research [9]. In addition, last phone durations have been taken into account.

The features mentioned above are used to derive binary indicator functions. For each word token w_i at position i textual and discretized durational information is taken into account from the area starting at word w_{i-2} until word w_{i+2} where pause p_i refers to the pause between words w_i and w_{i+1}, and duration d_i represents the duration of the last phone of w_i. Word based features include unigram, bigram, and trigram features, such as (w_{i-2}) or (w_{i-1}, w_i, w_{i+1}). For pauses and last phone durations, unigram and bigram features are composed. As found in [9] long pauses between two consecutive words are common both between two DAs and around hesitations. Therefore, combined word-pause features and word-phone duration features are added to the set of indicator functions, for example (w_i, p_i), or (w_i, d_i).

Finally, we exploit the fact that CRF are capable of learning sequences. For this, label bigram features are included. These features model the transition between consecutive word labels produced by the CRF model taking into account that not all label sequences are equally probable. Based on the presence of the label bigram feature we expect that coding schemes EI, BI, and BIE should exhibit better recognition performance than schemes E or B.

Reference	|S	S|Q	Q	Q	Q|D	D	D|B|			
System	|S|S|Q	Q	Q|S	S	S	S|B|				
Words	E	E	E	E	E	E	E	E	E	C
DAs	E	|	E		|	E	|C|			

Metric	Counts	Reference	Rate
F Measure	$2\,RP/(R+P)$	R, P	22.2
Recall R	1 correct DA	4 DAs	25.0
Precision P	1 correct DA	5 DAs	20.0
Strict	9 match errors	10 words	90.0
DER	3 match errors	4 DAs	75.0

Figure 2: Performance metrics for joint segmentation and classification of dialog acts.

2.3. Performance Metrics

To assess the performance of joint segmentation or classification of DAs we define the *F Measure* for the joint task based on *Precision* and *Recall* of dialog acts. A DA is considered to be correctly recognized only if both its type label matches the gold standard and it is correctly segmented according to the gold standard at the same time. For comparison to previous work we also provide results for the *Strict* metric [10], the *DER* metric [9], and the *NIST-SU* segmentation metric [14] (see Fig 2 for some examples). Note that using these definitions the *DER* metric is identical with $1-$ *Recall*.

3. Experiments and Discussion

To validate the approach based on CRF and the proposed coding schemes experiments have been conducted with the ICSI (MRDA) corpus applying two different classmaps under both reference and speech-to-text (STT) conditions[1]. The first classmap, *Map 01*, groups the 2083 unique DA labels found in the MRDA corpus into the five DA types statement, question, backchannel, floorgrabber, and disruption. The second classmap used here, *Map 05b*, preserves the most detail in the DA labels out of all the classmaps provided with the MRDA corpus. It further breaks down some groups of Map 01 into more specific DA types. For example, the question group is differentiated into more specific types, such as yes/no-questions, wh-questions, or-questions. From the statement group more detailed DA types such as responses to questions (rejection, acceptance, and uncertain) and action motivators (combining suggestions, commands, and commitments) are split away[2].

The experimental setup regarding the split of meetings to be used for training, validation, and tests was exactly set as in [8, 9, 10, 12], the development set was used for the selection of the feature sets and the initial CRF parameters. For all experiments the CRF++ implementation [16] was used with default settings

[1] For the STT output the recognizer described in [15] was used achieving a word error rate of 35.4% on the close talking microphones.

[2] Consult the documentation provided with the MRDA corpus for details.

DA Coding	F Measure	Strict	DER	NIST-SU
E	52.3	61.9	52.0	32.2
B	50.1	63.7	54.7	32.5
EI	52.6	61.6	51.6	32.4
BI	52.4	61.3	52.0	32.0
BIE	53.8	60.0	50.4	31.1

Table 1: Test set results for investigated DA coding schemes under reference condition.

applied (unless explicitly mentioned). In the text below only results for the test sets are listed, as all the findings are equally valid for both development and test sets.

The goal of the first experiment was to set base line results for the CRF approach investigated in this paper. For this, coding scheme E, classmap *Map 01*, and reference transcriptions were chosen to be able to compare results against those of previous work. Test set results for this task are provided in the first line of Table 1 as well as the upper part of Table 2. As can be observed from the results listed in Table 2 the approach presented here significantly outperforms the figures published in [7, 9, 10] for both the Strict and the NIST-SU metric. Computing the F measure for the five individual DA types shows great differences in the performance. The DA types recognized best are backchannels at an F measure of 83.3, followed by floorgrabbers at 52.7, and statements at 50.6. The poorest results are obtained for questions at an F measure of 32.6 while disruptions reach 36.8. Although the overall performance for the system applying Coding B as reported in Table 1 is clearly lower than the performance obtained with Coding E this observation does not hold if the performance for individual DA types is compared between Coding B and Coding E. A substancial F measure gain from 32.6 to 40.7 is obtained for questions while disruption performance drops from 36.8 down to 19.8. Considering that questions very frequently start with specific words such as *right*, *is*, *how* and disruptions are much more characterized by their unfinished endings than by their first word this result well matches the expectations.

Results shown in Table 1 for Scheme EI, and BI show improvements over the corresponding figures for Schemes E, and B respectively. The fact that the performance gain is larger for Scheme BI can be explained by the F measure increase from 19.8 to 33.5 for disruptions which seem to profit most of the DA specific non-boundary tokens. Finally, as hypothesized, scheme BIE combining the advantages of scheme B (good question detection) and scheme E (disruption detection) results in a significant performance improvement over all other coding schemes investigated. Again, backchannels and floor grabbers are recognized best with F measures of 83.4, and 55.2 followed by statements at 51.6 and questions at 45.8 respectively. The F measure of 37.2 for disruptions indicates that this group of DAs is the hardest to detect reliably for classmap *Map 01* which confirms earlier results for an approach based on graph search [8].

To further validate the CRF based approach the next experiments focus on the a more fine-grained classmap *Map 05b* that differentiates 17 DA types instead of the five DA types defined by *Map 01* for the MRDA corpus. The interest for this classmap also arises from the domain of action item identification where it has been found that fine-grained DA types may help to improve performance [2, 3] in contrast to the use of the classmap *Map 01* which did not prove to be helpful [17].

Results for classmap *Map 05b* under reference conditions are listed for schemes E and BIE in Table 2. Using Scheme E the CRF has to solve a 18 class problem while for Scheme BIE 86 classes are to be differentiated[3]. As expected, a lower performance is observed for the joint task caused by the more than three-fold increase in the number of DA types. The significantly better results for coding scheme BIE than scheme E are consistent with the related findings for *Map 01*. If F measures are computed per DA type the same initial ranking as for *Map 01* is obtained. Backchannels perform best at an F measure of 73.9 followed by floorgrabbers at 53.5, and statements at 41.7 when using scheme BIE. Statements are followed by wh-questions and yes/no-questions (representing over 80% of the question group) with F measures of 40.1, and 37.8 respectively. From the response group rejections are detected most reliably at an F measure of 37.1, followed by DAs of type uncertain at 34.8. With an F measure of 25.0 DAs related to acceptance are detected at a substantially lower level. Acceptance DAs are frequently mistaken with backchannels, floor grabbers, or statements which follows from the proposed approach that (for now) does not take into account interactions between speakers and the fact that acceptance is per definition related to a question/request posed by a second speaker. As for *Map 01*, the detection of disruptions only reaches a relatively low F measure of 36.3. Finally, the detection of action motivator DAs that combines commands, suggestions, and commitments into a single group seems to be particularly challenging. For this group an F measure of 16.0 is achieved where action motivators are most frequently mistaken for statements. As this group largely consists of suggestions stretching over many words the current feature set is most likely not designed to clearly separate statements from action motivators. For further processing such as action item identification substantially higher action motivator detection rates will most likely be required.

Results for STT conditions are listed in the lower half of Table 2 for both classmaps as well as coding schemes E and BIE. Note that a direct comparison of results to previous work [9, 10] is difficult as these results are based on a previous version of the STT module with a lower recognition performance. A detailed analysis of results for both reference and STT conditions leads to the observation that all main findings for reference conditions also hold for the STT conditions. Specifically, coding scheme BIE outperforms scheme E significantly due to the better question identification with an F measure of 31.0 for scheme BIE, and 18.9 for scheme E. Interestingly, the performance drop from reference conditions to STT conditions for the joint task is larger for the simpler DA classmap *Map 01* than the more fine-grained *Map 05b*. This finding holds for both different coding schemes as well as performance metrics applied.

4. Conclusion and Outlook

We investigated the use of CRFs for joint segmentation and classification of dialog acts. Based on extensive experimental validation on both coarse and fine-grained tagsets we conclude that the proposed approach is attractive for both the achieved performance as well its conceptual simplicity. The application of a coding scheme that preserves contextual information for the first word, the last word, as well as words inside of dialog

[3] To keep the model size for this 86 class configuration tractable the used features were restricted to those that occured at least 15 times in the training data.

Conditions	Classmap	Classes	System	DA Coding	F Measure	Strict	DER	NIST-SU
Reference	Map 01	5	[10]	n/a	n/a	64.4	54.4	34.5
		5	[9]	E	n/a	62.8	51.0	34.8
		5	[7]	n/a	n/a	62.7	47.4	32.0
		5	CRF	E	52.3	61.9	52.0	32.2
		5	CRF	BIE	53.8	60.0	50.4	31.1
	Map 05b	17	CRF	E	45.3	66.7	55.3	32.2
		17	CRF	BIE	46.9	64.7	53.7	31.5
STT	Map 01	5	[10]	n/a	n/a	75.4	64.3	45.5
		5	[9]	E	n/a	73.6	62.6	44.6
		5	CRF	E	43.7	70.3	60.9	40.5
		5	CRF	BIE	44.5	69.9	59.9	40.6
	Map 05b	17	CRF	E	37.9	73.7	63.4	40.7
		17	CRF	BIE	39.2	72.4	62.0	39.6

Table 2: Performance comparison of the CRF approach investigated in this paper with systems published previously for the same experimental setup. Test set results are provided under both reference and STT conditions using classmap *Map 01* representing 5 distinct dialog act types and *Map 05b* representing 17 dialog act types, respectively. For [7] the results referring to the *hybrid* system are reported.

acts substantially boosts question detection compared to previous work.

Future work is possible in many areas. Results for the fine-grained DA tagset indicate that the automatic detection of responses and action motivators requires further attention. Both groups of DAs might profit from taking into account features related to speaker interactions. Alternatively to the use of conventional CRF studied here semi-Markov CRF could prove an interesting alternative as they are capable to integrate features defined on complete dialog acts. Finally, it would be interesting to integrate joint segmentation and classification of DAs with related fields such as speaker segmentation or action item identification.

5. References

[1] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Jiekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting speech recognition and understanding system," in *Proc. SLT Workshop*, Goa, India, 2008, pp. 69–72.

[2] W. Morgan, P.-C. Chang, S. Gupta, and J. M. Brenier, "Automatically detecting action items in audio meeting recordings," in *7th SIGdial Workshop on Discourse and Dialogue*, Sidney, Australia, 2006, pp. 69–103.

[3] F. Yang, G. Tur, and E. Shriberg, "Exploiting dialog act tagging and prosodic information for action item identification," in *Proc. ICASSP*, Las Vegas, USA, 2008, pp. 4941–4944.

[4] A. H. Buist, W. Kraaij, and S. Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Meeting of Computational Linguistics in the Netherlands*, Leiden, The Netherlands, 2004, pp. 28–35.

[5] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. 5th Europ. Conf. on Speech, Communication, and Technology*, Rhodes, Greece, 1997, vol. 1, pp. 207–210.

[6] E. Shriberg et al., "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGDIAL*, Cambridge, USA, 2004, pp. 97–100.

[7] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1303–1314, 2008.

[8] M. Zimmermann et al., "A* based joint segmentation and classification of dialog acts in multi-party meetings," in *Proc. 9th ASRU*, San Juan, Puerto Rico, 2005, pp. 215–219.

[9] M. Zimmermann, A. Stolcke, and E. Shriberg, "Joint segmentation and classification of dialog acts in multi-party meetings," in *Proc. 31st ICASSP*, Toulouse, France, 2006, vol. 1, pp. 581–584.

[10] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, Philadelphia, USA, 2005, vol. 1, pp. 1061–1064.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, Williamstown, USA, 2001, pp. 282–289.

[12] M. Zimmermann et al., "Text based dialog act classification for multiparty meetings," in *Machine Learning for Multimodal Interaction: 3rd International Workshop, MLMI 2006*, Washington, USA, 2006.

[13] S. Sarawagi and W. Cohen, "Semi-Markov conditional random fields for information extraction," in *ICML*, Banff, Canada, 2004.

[14] NIST website, "RT-03 fall rich transcription," http://www.nist.gov/speech/tests/rt/rt2003/fall/, 2003.

[15] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Machine Learning for Multimodal Interaction: 2nd International Workshop, MLMI 2005*, S. Renals and S. Bengio, Eds., pp. 463–475. LNCS 3869, Springer, 2006.

[16] T. Kudo, "CRF++: Yet another CRF tool kit," http://crfpp.sourceforge.net/, 2005.

[17] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multiparty dialgue," in *8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, 2007.