DIALOGUE-ACT TAGGING USING SMART FEATURE SELECTION; RESULTS ON MULTIPLE CORPORA

Daan Verbree, Rutger Rienks, Dirk Heylen

Human Media Interaction Lab, University of Twente, Enschede, the Netherlands

ABSTRACT

This paper presents an overview of our on-going work on dialogue-act classification. Results are presented on the ICSI, Switchboard, and on a selection of the AMI corpus, setting a baseline for forthcoming research. For these corpora the best accuracy scores obtained are 89.27%, 65.68% and 59.76%, respectively. We introduce a smart compression technique for feature selection and compare the performance from a subset of the AMI transcriptions with AMI-ASR output for the same subset.

1. INTRODUCTION

The topic of automatic Dialogue Act classification has received a fair amount of attention in the past years (1; 2; 3) (see also Table 1). A variety of methods have been tested on various corpora using different dialogue act classes. This can make the comparison between different methods rather difficult. It is well known that the words and phrases in DA's are the strongest cues to their identity (1). When looking at current state-of-the-art DA tagging, we may conclude that experiments that are easily and unambiguously replicable and that compare the performances on different corpora have not yet been conducted. This paper describes the first session of a series of experiments that tries to adhere to these issues. We next describe the three corpora that we used, provide an overview of previous work on these corpora, explain our approach and then compare our performances on the three corpora with known results. Finally, we present first results on DA classification on ASR output, instead of on manual transcriptions.

2. VARIOUS CORPORA

For our DA tagging experiments we have used three different corpora, each with their own tagset: the ICSI Meeting Corpus (4), the Switchboard Corpus (5) and part of the AMI corpus (6). We briefly describe each of these in turn.

The ICSI Meeting Corpus includes 75 naturally occurring meetings containing roughly 72 hours of multi-talk speech data and associated human generated word-level transcripts. It was hand-annotated for dialog acts as described in (7; 3) using the Meeting Recorder Dialog Act tagset (MRDA). The MRDA scheme has 11 general tags and 39 specific tags. Each annotation requires one general tag and a variable number of specific tags. The MRDA scheme proposes several classmaps as well in which several tags are grouped together. For our experiments on the ICSI corpus we have used the widely applied classmap that maps the DA's onto 5 distinct classes: statements (S), questions (Q), backchannels (B), fillers (F) and disruptions (D). Utterances that have not been annotated are labelled (Z).

The ICSI Corpus comes along with a proposed train/test split. This split consists of 51 meetings (almost 80.000 utterances) which can be used for training, 11 meetings (about 13.500 utterances) for development, and 11 meetings (over 15.000 utterances) for evaluation. This split leaves out 2 of the 75 meetings. These are excluded because of their different nature.

The Switchboard Corpus is a corpus of conversational speech by telephone. For our experiments, we used the same subset of the corpus as (8). The subset consists of over 210,000 utterances grouped in 1,155 conversations. Dialogue act annotations based on the SWBD-DAMSL tagset are available for all of these conversations (9). Similar to (8), we used the clustered tagset containing 42, out of the original 220 DA-labels.

The AMI Corpus is a collection of over 100 hours of four person project meetings. All meetings are in English. However a large proportion of speakers are non-native English speakers. Amongst a lot of signals, the transcriptions of all meetings are available as well as several *layers* of annotations. Since not all the dialog annotations of the meetings were available at the time we ran the experiments, we used a subset of 80 meetings[1]. Our collection comprises about 50.000 utterances. The AMI DA tagset has 15 tags : Backchannel, Stall, Fragment, Inform, Elicit Inform, Suggest, Offer, Elicit Offer Or Suggestion, Assess, Elicit Assessment, Be Positive, Be Negative, Comment About Understanding, Elicit Comment About Understanding, and Other.

3. PREVIOUS WORK

3.1. Baseline

A baseline used for comparing the accuracy of classification results, is the majority class baseline. We choose however, to compare the performance with the performances achieved using the set of manually acquired cue phrases, known as the LIT set, as proposed by Samuel (24). This set contains 687 different cue phrases that have been assembled from several papers, dissertations and books. Table 2 gives an overview of the baselines computed for the three corpora. For the Switchboard corpus we were unable to compute the LIT set performance on a machine with 2048 MB internal memory.

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-196).

[1]These were: ES2002ACD; ES2003BCD; ES2006; ES2007; ES2008; ES2009; ES2010; ES2011ABC; ES2012CD; ES2014ABC; ES2015; ES2016; IS1000A; IS001; IS1003; IS1004; IS1005ABC; IS1006ABD; IS1008; IS1009; TS1003ABC; TS1004ABC; TS1005; TS1007A.

Feature / Article	(10)	(11)	(12)	(2)	(13)	(14)	(15)	(16)	(17)	(18)	(1)	(19)	(20)	(21)	(22)	(23)
Sentence length	X								X							
First two words	X	X														
Last two words	X															
First word of next segment	X															
Speaker		X														
Number of utterances		X														
Prosodic			X		X						X			X		
Bigrams of words in segment				X												
(Correct) last 10 DA's					X											
Words in last 10 DA's					X											
Utterance type						X										
Presence/absence 'Wh'-words						X										
Subject Type						X										
Specific cue phrases						X					X				X	
First verb type						X										
Second verb type						X										
Question mark						X										
Polygrams of words in segment							X							X		
Ngrams of words in segment								X		X		X	X			X
Ngrams of previous DA's								X		X			X			
Specific patterns									X							
Previous DA									X							
Next DA									X							
Grammar pattern									X		X					

Table 1. Features used for DA-classification in different studies

Corpus	Majority Class	LIT set
ICSI	55.46	63.57
Switchboard	33.73	uncomputable
AMI	28.69	44.24

Table 2. Baselines for the different corporas

3.2. Performances

Most studies on dialogue act classification have used one of the corpora that we use in our experiments. We present the best results obtained, as far as we know, for the ICSI and Switchboard corpus. As the AMI Corpus is a new corpus no previous DA tagging results have been published yet.

The best performance for DA classification on the ICSI corpus currently is 81.18%, as reported in (10). The classification obtained, was performed on the same train/test split and using nearly (omitting 'Z' class) the same classmap as mentioned before in Section 2. This results in a classification task with 5 distinct classes.

Previous research on the DA classification of the Switchboard Corpus has been reported by Rotaru et al. (2) and Stolcke et al. (8). Stolcke obtains a 71% accuracy using a trigram word model whereas Rotaru achieves 72%. Unfortunately both use different fixed train/test splits, without mentioning which split has been used for these results. Furthermore the description of the methods used are not very clear. Stolcke uses trigrams and Rotaru uses bigrams as one of the features, but in both cases it is unclear if n-gram selection methods have been used. Another difference between the two studies is that Rotaru included the utterances labelled with the '+' DA-tag, whereas Stolcke excluded these. Based on these descriptions it is not possible to reproduce nor to compare their results.

4. OUR APPROACH

We tried to devise an experiment that was both replicable. Our used feature-set contained the following categories:

?/OR: Whenever a question mark is present, the number of times the word *or* appears is counted and used as a feature.[2]

Length: The length (number of words) of each segment.

Last Labels: A bi-gram of the previous two labels.

N-grams (compressed): At first, all bi-, tri- and quadri-grams of words were computed for all tagged utterances. Then we applied the n-gram selection method which selects the Top-N most predictive n-grams using the ranker explained in (25), which in turn was inspired by (26). This was done for the Top-N of the uni, bi-, tri- and quadri-grams (order specific) and the Top-N of the merge of all n-grams (non order specific) for each class. These remaining n-grams are used for calculation of the feature values for each speech act. The final step awards a number of points based on the match with the preselected n-grams (See (25) for more details). Instead of a 'presence' value for each individual pre-selected n-gram, the information is now *compressed* into just as many feature values as there are class labels (for each type of N-gram). This enables e.g. performance computations on the huge Switchboard corpus.

POS-N-grams (compressed) The POS-N-gram features were computed in a similar fashion as described for the word n-grams. POS-tag features have been scarcely used for DA-classification, if

[2]Until now our classification is based on transcriptions in which the question mark is available, but eventually we aim to base our classification on ASR-output in which this feature might not be available anymore.

at al. (27) have used them for back-channel classification, but none of the papers presented in Table 1 mentioned its usage.

The most significant difference with our approach in comparison to earlier approaches is the use of a *compressed* feature set for the N-gram and POS-N-gram features. This technique enabled us, in contrast to e.g. (10) to make use of each word of the utterance and unlike e.g. (20) we did not end up with an extremely large feature set. Table 4 shows that there is hardly any difference between the results of the compressed (C) version and the uncompressed (I) version. The other abbreviations used in the remainder of this paper to describe our features are listed in Table 3.

Abbreviation	Feature
L	Length
P	Uni-, bi-, tri- and quadrigrams of POS-tags
W	Uni-, bi-, tri- and quadrigrams of words
QMT	Question mark token
ORT	'OR' token
LL	Last Label
NOS	Non-Order Specific
OS	Order Specific
C	Compressed
I	Individual
T10	Top 10

Table 3. Features and their abbreviations

5. RESULTS

All results presented were obtained by using the J48 classifier using the default settings as available in Weka (28). J48 was chosen after a careful selection of classifiers in which both performance and computational time were taken into account. (25). For computation of all our results, 10-fold cross validation was used.

To compare the effect of the compression, the results of the compressed set are (when computationally possible) compared with the feature set containing the combination of all the individual (Top-N) Ngrams.

For the classification results on the ICSI corpus we used the train/test split provided. Contrary to Ang & Shriberg (10), we did not exclude utterances of the class 'Z'. The results obtained in our experiments are depicted in Table 4.

Feature set and parameters chosen	Performance
L_P_W (OS) (C T10)	87.84
L_P_W (OS) (I T10)	87.97
QMT_ORT_L_P_W (OS) (C T10)	87.82
QMT_ORT_L_P_W (OS) (I T10)	87.98
QMT_ORT_L_LL_P_W (OS) (C T10)	89.13
QMT_ORT_L_LL_P_W (OS) (I T10)	89.27

Table 4. DA-classification results on the ICSI Meeting corpus

On the Swichboard corpus we used the 42 DA's from the classmap (similar to Stolcke et al.) and also included the *'+'* DA by using it on its own. (The *'%-'* DA was mapped on the *'%'* DA.) Note that this makes the results not *directly* comparable to those of Stolcke and Rotaru. To overcome this, we have also performed an experiment in which we discarded the utterances of the class *'+'*, resulting in an accuracy of 70.26%.

Our performances on the Switchboard corpus is shown in Table 5. Similar to the inability to compute the performance of the LIT set on the Swichboard corpus, we were unable to compute results for the classifiers using individual n-grams. This is mainly due to the amount of distinct DA tags used and the size of the corpus. It should be said that our computations on the Switchboard Corpus still required a huge amount of computing power, even for the compressed feature set. The process of part-of-speech tagging, ngramming and classifying all 10 folds created out of the 210,000 utterances available in the Switchboard Corpus took about 3 days for each classifier-setting.

Feature set and parameters chosen	Performance
L_P_W (OS) (C T10)	60.57
QMT_ORT_L_P_W (OS) (C T10)	60.22
QMT_ORT_L_LL_P_W (OS) (C T10)	65.68

Table 5. DA-classification results on the switchboard corpus

For the AMI corpus we used the proposed train and test as described in AMI D5.2 (29). Table 6 lists all the obtained performances. The best performance achieved is 59.76%. For this an additional preprocessing step was performed where for each type of N-gram the values for the class labels were transformed into a binary one. The class label with the highest value was given value one, whereas the other class labels all were set to zero. Furthermore the top 300 N-grams were used instead the top 10.

Feature set and parameters chosen	10 fold
L_P_W (OS) (C T10)	53.94
QMT_ORT_L_LL_P_W (OS) (C T10)	55.54
QMT_ORT_L_LL_P_W (OS) (C T300)*	59.76

Table 6. DA-classification results on the AMI corpus. The $*$ indicates that the extra preprocessing step was performed.

6. RESULTS ON ASR

In order to move one more step in the direction of fully automatic DA classification, we have performed a DA classification experiment on 11 AMI meetings[3] from which the ASR output was available. The DA-labels used originated from the annotations performed on the manual transcripts. The resulting corpus consisted of 8374 utterances. The experiments were run in a 10 fold cross-validation setup. Table 7 shows the performances obtained for this ASR corpus in comparison to the results using the manual transcriptions for the same set of meetings.

The decrease in performance, which was expected, can in potential be blamed to the word error rate of the speech recognizer.

[3]ES2002ACD; ES2009ABCD; IS1009ABCD

Feature set and parameters chosen	ASR	Manual
L_P_W (OS) (C T10)	37.05	51.29
L_P_W (OS) (I T10)	40.26	56.55
QMT_ORT_L_P_W (OS)(C T10)	37.43	53.74
QMT_ORT_L_P_W (OS) (I T10)	40.05	56.89

Table 7. DA-classification results on Manual and ASR transcriptions

However, also the smaller corpus size may have played an important role, as some of the features that were used become more useful on a larger data-set (c.f. language models). We need to wait until more data is available before we can further address this issue.

7. DISCUSSION

Closer analysis of the QMT_ORT_L_LL_P_W (Order Specific) (Individual Top 10) classifications on the ICSI corpus (see its confusion matrix in table 8) shows one of the most interesting challenges for future work. It appears that a lot of *backchannels* are misclassified as *statements*. Analyzing the ngrams selected which should cue for *backchannels* it appears that, even if an ngram is among the best-cueing ngrams for a specific class it might even cue more for another class. This phenomena was observed for all three corpora examined.

a	b	c	d	e	f	$<--$ classified as
1558	1	18	0	384	0	a = B
55	1869	131	125	60	4	b = D
133	100	990	0	91	3	c = F
0	12	0	1095	4	4	d = Q
471	2	14	19	8057	6	e = S
5	4	3	0	3	177	f = Z

Table 8. Confusion matrix of QMT_ORT_L_LL_P_W (Order Specific) (Individual Top 10) ICSI setting

As a result of this, also reflected in Table 8, the more frequently occurring classes also have a larger chance of being classified correctly. Normalization in a preprocessing phase could potentially overcome this.

A result of our multi-corpus approach is that it brings us in a position where we are able to investigate the impact of various corpus types (Meetings, v.s. Telephone conversations) on the performance. Readers should note that, since our classification results are better than Shribergs' and equivalent to Stolcke's, this does not e.g. legitimates to infer that Stolcke's classification performance on Switchboard outperforms Shriberg's performance on the ICSI corpus. One cannot say this because features that work well on one corpus could work even better, or worse on a different corpus.

8. CONCLUSIONS

In this paper we presented a method of DA tagging using a *compressed* feature set that apart from using words also used the more general part-of-speech-level of a sentence. Results on different corpora show a major improvement over the majority class baseline as well as over the LIT set baseline. Furthermore our classification outperforms earlier results obtained on the ICSI set sets a inter-corpora standard for the Switchboard and AMI corpus, using a replicable 10 fold cross-validation approach. The results on the ASR output show a large decrease in performance.

References

[1] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts," in *Discourse Relations and Discourse Markers: Proceedings of the Conference*, Manfred Stede, Leo Wanner, and Eduard Hovy, Eds., pp. 114–120. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[2] M. Rotaru, "Dialog act tagging using memory-based learning," Tech. Rep., University of Pittsburgh, spring 2002, Term project in Dialogue-Systems class.

[3] E. Shriberg, R. Dhillon, and S. Bhagat et al., "The icsi meeting recorder dialog act (mrda) corpus," in *Proc. HLT-NAACL SIGDIAL Workshop*, 2004.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, pp. 364–367.

[5] J Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development.," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, San Francisco, March 1992, vol. 1, pp. 517–520.

[6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meetings corpus," in *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*, 2005, AMI-108.

[7] R. Dhillon, S. Bhagat, H Carvey, and E. Shriberg, "Meeting recorder project: Dialogue act labeling guide," Tech. Rep., ICSI Speech Group, Berkeley, USA, 2003.

[8] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[9] D. Jurafsky, E. Shriberg, and D. Biaska, "Switchboard SWBD-DAMSL shallow-discourse-function annotation (coders manual, draft 13)," Tech. Rep., Univ. of Colorado, Inst. of Cognitive Science, 1997.

[10] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proceedings of the 30th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2005.

[11] S. Rosset and L. Lamel, "Automatic detection of dialog acts based on multi-level information," in *icslp*, Jeju Island, October 2004, pp. 540–543.

[12] R. Fernandez and R.W. Picard, "Dialog act classification from prosodic features using support vector machines," in *Proceedings of speech prosody 2002*, April 2002.

[13] P. Lendvai, A. van den Bosch, and E. Krahmer, "Machine learning for shallow interpretation of user utterances in spoken dialogue systems," in *Proceedings of EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management*, 2003, pp. 69–78.

[14] T. Andernach, "A machine learning approach to the classification of dialogue utterances," *Computing Research Repository*, vol. july, 1996.

[15] N. Reithinger and M. Klesen, "Dialogue act classification using language models," in *Proceedings of EuroSpeech-97*, 1997, pp. 2235–2238.

[16] A. Venkataraman, A. Stolcke, and E. Shirberg, "Automatic dialog act labeling with minimal supervision," in *Proceedings of the 9th Australian International Conference on Speech Science & Technology*, December 2002.

[17] S. Keizer and R. op den Akker, "Dialogue act recognition under uncertainty using bayesian networks," *Natural Language Engineering*, vol. 1, pp. 1–30, 2005.

[18] A. Venkataraman, Y. Liu, E. Shriberg, and Stolcke A., "Does active learning help automatic dialog act taggin in meeting data," in *Proc. Eurospeech*, 2005.

[19] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," in *Proceedings of the MLMI*.

[20] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "A* based joint segmentation and classification of dialog acts in multiparty meetings," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, november 2005.

[21] V. Warnke, R. Kompe, H. Niemann, and E Noth, "Integrated dialog act segmentation and classification using prosodic features and language models," September 1997, pp. 207–210, Eurospeech.

[22] S. Katrenko, "Textual data categorization: back to the phrase-based representation," in *Proceedings in 2nd International IEEE Conference "Intelligent systems", Vol. III*, June 2004, pp. 64–67.

[23] N. Webb, M. Hepple, and Y. Wilks, "Dialogue act classification based on intra-utterance features," in *Proceedings of the AAAI 05*.

[24] K. Samuel, *Discourse Learning: An Investigation of Dialogue Act Tagging using Transformation-Based Learning.*, Ph.D. thesis, Department of Computer and Information Sciences. University of Delaware. Newark, Delaware., 2000.

[25] A.T. Verbree, "On the structuring of discussion transcripts based on utterances automatically classified," M.S. thesis, University of Twente, 2006.

[26] K. Samuel, S. Carberry, and K. Vijay-Shanker, "Automatically selecting useful phrases for dialogue act tagging," *The Computing Research Repository*, June 1999.

[27] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 2003, pp. 51–58, Association for Computational Linguistics.

[28] I. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, 2000.

[29] J. Alexandersson, R. Op den Akker, T. Becker, Y. Goto, S. Hsueh, N. Jovanovic, T. Kleinbauer, W. Kraaij, I. McCowan, J. Moore, B. Peskin, R. Rienks, A. Renalds, S. Alesandro Viciarelli, and W. Xu, "Implementation and evaluation results (ami d.5.2)," Tech. Rep., AMI Project, July 2006.