Creating Annotated Dialogue Resources: Cross-Domain Dialogue Act Classification

Dilafruz Amanova[1], **Volha Petukhova**[2], **Dietrich Klakow**[2]
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Saarland University, Spoken Language Systems, Saarbrücken, Germany
[1]dilafruz@mpi-inf.mpg.de;
[2]{v.petukhova,dietrich.klakow}@lsv.uni-saarland.de

Abstract

This paper describes a method to automatically create dialogue resources annotated with dialogue act information by reusing existing dialogue corpora. Numerous dialogue corpora are available for research purposes and many of them are annotated with dialogue act information that captures the intentions encoded in user utterances. Annotated dialogue resources, however, differ in various respects: data collection settings and modalities used, dialogue task domains and scenarios (if any) underlying the collection, number and roles of dialogue participants involved and dialogue act annotation schemes applied. The presented study encompasses three phases of data-driven investigation. We, first, assess the importance of various types of features and their combinations for effective cross-domain dialogue act classification. Second, we establish the best predictive model comparing various cross-corpora training settings. Finally, we specify models adaptation procedures and explore late fusion approaches to optimize the overall classification decision taking process. The proposed methodology accounts for empirically motivated and technically sound classification procedures that may reduce annotation and training costs significantly.

Keywords: cross-domain dialogue act classification, ISO 24617-2 annotated resources, dialogue act model adaptation, meta-classification

1. Introduction

Corpora annotated with semantic information gained lots of researchers' and practitioners' interest and are acknowledged to be important for a wide range of linguistic applications. In the dialogue research area, it became common to annotate dialogue corpus data with dialogue act information. Dialogue corpus annotations may serve various purposes. Annotated data is used for a systematic analysis of a variety of dialogue phenomena, such as turn-taking, feedback, and recurring structural patterns. Corpus data annotated with dialogue act information are also used to train machine learning algorithms for the automatic recognition and prediction of dialogue acts as a part of a human-machine dialogue system.

The dialogue research community still does not have large amounts of annotated dialogue data at its disposal as compared to other linguistic communities. Although a lot of work has been done in this direction, e.g. the HCRC MapTask corpus (Carletta et al., 1996), the AMI[1] and ICSI-MRDA (Dhillon et al., 2004) meeting corpora, the Switchboard-DAMSL (Jurafsky, 1997) and Coconut (Di Eugenio et al., 1998) corpora, just to name few. The available resources however are only partly compatible with each other. Annotations are based on schemes developed for analysis and modelling corpus- and domain-specific dialogue behaviour, and annotation formats vary a lot ranging from plain text to XML based representations, e.g. `SGML` and `NXT`. Annotations differ with respect to the range of phenomena they cover, their granularity level of defined concepts, segmentation principles and their theory dependencies. Thus, such corpora are not easy to re-use for purposes and apply to domains other than they were originally developed for. For cross-domain analysis it is important to have interoperable annotated resources. One way to achieve interoperability is to create resources by annotating new data collections using existing standards or at least the subset of their main concepts with a well-defined relation to those of the standard. The ISO 24617-2 dialogue act annotation scheme (ISO, 2012) serves these purposes. The ISO dialogue act annotation scheme has been already deployed in some dialogue projects, e.g. the ToMA project (Blache et al., 2009) where the Corpus of Interactional Data (CID) was labelled according ISO 24617-2, and the DBox project dialogue gaming data (Petukhova et al., 2014). Another possibility is to convert the annotations in existing corpora to annotations that are compatible with the ISO standard. For example, this approach has been applied to the SWBD-DAMSL annotations in the Switchboard corpus (Fang et al., 2012; Bunt et al., 2013). The third way of achieving interoperability is to manually map various annotation schemes' concepts to those of ISO 24617-2 (Petukhova, 2011).

All described approaches lead to the creation of annotated interoperable resources which become rather expensive. The analyses reported in (Petukhova and Bunt, 2007) showed that the ratio of annotation time to real dialogue time was approximately 19:1 when coding by expert annotators. Moreover, time and resources are required to train annotators and evaluate their work. In this paper we propose a method how to reduce dialogue act annotation costs by applying prediction models obtained by training multiple machine learning classifiers on available annotated dialogue resources, and further adapting these models to dialogue data of various domains.

We, first, discuss work that has been performed on dialogue act classification and multi-corpus dialogue act recognition. Further, we present our training data that we created using the method proposed by Petukhova et al. (2014), namely

[1] http://groups.inf.ed.ac.uk/ami/corpus/

by accessing existing annotated corpora through a mapping from ISO 24617-2 concepts to those of the annotation scheme used in the corpus. Subsequently, we describe the experimental set up, provide details on different training and feature selection settings, and outline results. Finally, we define dialogue act classification model adaptation task and propose two potential system designs for automatic creation of the dialogue act annotated resources. We wrap up the paper by summarizing obtained results and outlining future research.

2. Related Work

The recognition of the intentions encoded in user utterances is one of the most important aspects of language understanding for a dialogue system. Various machine learning techniques have been applied successfully to natural-language based dialogue analysis. For example, Hidden Markov Models (HMM) have been applied to dialogue act classification in the Switchboard corpus (Stolcke et al., 2000), achieving a tagging accuracy of 71% on word transcripts. Another approach that has been applied to dialogue act recognition, by Samuel et al. (1998), uses transformation-based learning. They achieved an average tagging accuracy of 75.12% for the Verbmobil corpus. Keizer (2003) used Bayesian Networks applying a slightly modified version of DAMSL with an accuracy of 88% for backward-looking functions and 73% for forward-looking functions in the SCHISMA corpus.[2] Lendvai et al. (2004) adopted a memory-based approach, based on the k-nearest-neighbour algorithm, and report a tagging accuracy of 73.8% for the OVIS data, train information-seeking dialogues in Dutch.

Apart from using different techniques, these approaches also differ with respect to feature selection strategies. Some approaches rely solely on the wording of an input utterance, using n-gram models or cue-phrases, e.g. Reithinger (1997) and Webb et al. (2005). Others successfully integrate prosodic features that facilitate accurate dialogue act recognition, e.g. Shriberg et al. (1998); Jurafsky et al. (1998a); Fernandez and Picard (2002); Stolcke et al. (2000). Again others combine the predictions derived from the utterance and its context, e.g. Keizer (2003); Stolcke et al. (2000); Samuel et al. (1998); Lendvai et al. (2004)

In (Webb and Liu, 2008), authors carried out multi-corpora classification experiments based on purely intra-utterance features, principally involving word n-gram cue phrases. Automatically extracted cues from one corpus were applied to a new annotated data set, to determine the portability and generality of learned cues. It was shown that automatically acquired cues are general enough to serve as a cross-domain classification mechanism with an accuracy of 70.8%. Experiments were carried out on SWBD and the AMITIES GE corpora, where both corpora were annotated based on the DAMSL annotation scheme.

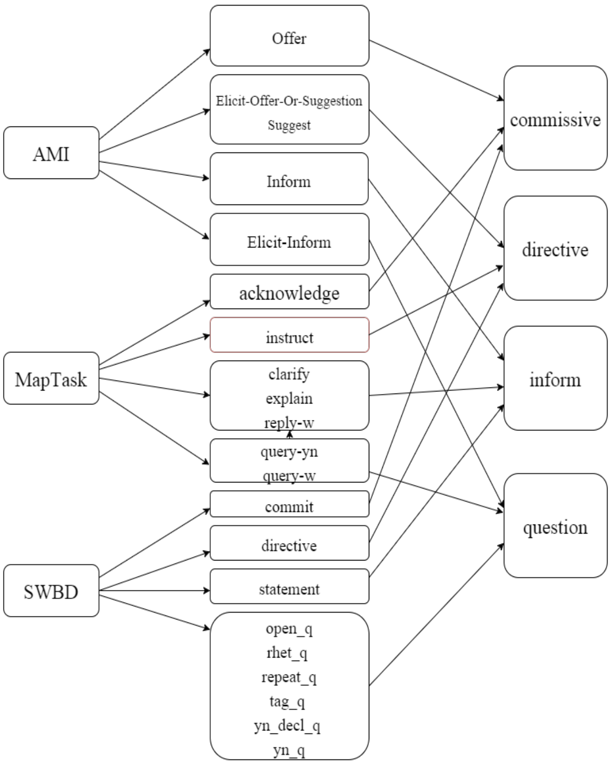


Figure 1: AMI, HCRC MapTask and SWBD-DAMSL dialogue act mapping to four ISO 24617-2 categories.

DA Type	AMI	MapTask	SWBD	Metalogue
Commissives	2.0	21.0	3.0	19.5
Directives	8.0	15.1	13.0	20.0
Inform	26.6	11.5	36.0	20.5
Question	3.4	17.0	4.0	20.0
Other tag	60.0	35.4	44.0	20.0

Table 1: Relative frequencies of dialogue acts categories in four different corpora. See also (Dielmann and Renals, 2008), (Surendran and Levow, 2006), and (Popescu-Belis, 2003)

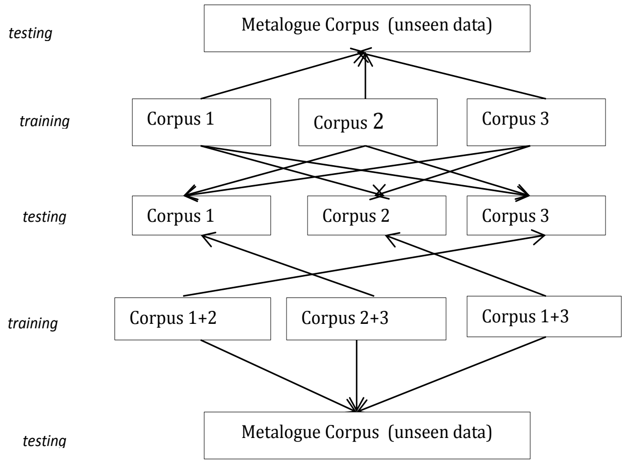
Webb et al. (2010) performed research on cross-domain dialogue act classification based on manually extracted cue phrases. For the experiments they used SWBD DAMSL and ICSI-MRDA corpora, which were annotated using variants of the DAMSL annotating scheme, achieving 72.34% accuracy.

The main goal of this study is to investigate to what extent multiple classifiers can handle different types of dialogue data collected for domains and tasks of various complexities, and annotated with principally different dialogue act annotation schemes.

3. Annotated Dialogue Data: Training and Test Sets

We considered three different corpora: AMI, HCRC MapTask and Switchboard. These corpora feature different types of interaction (two-party vs multi-party; face-to-face vs telephone conversations) and domains (meetings, instruction providing and free conversations). They are annotated using three different tag sets (AMI, HCRC Map-

[2] The SCHISMA corpus consists of 64 dialogues in Dutch collected in Wizard-of-Oz experiments, has keyboard-entered utterances within the information exchange and transaction task domain, where users are supposed to make inquiries about theatre performances scheduled and make ticket reservations.



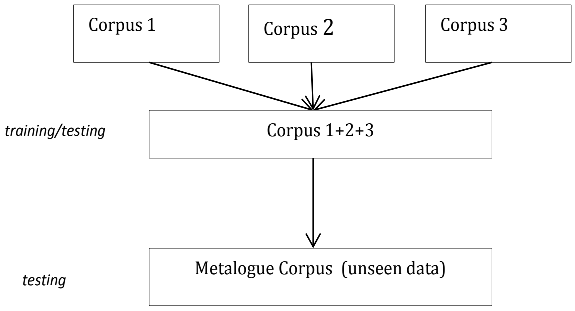


Figure 2: Cross-domain dialogue act classification: training/testing design.

Task and SWBD-DAMSL schemes[3]). These corpora form important dialogue community resources and they are the biggest annotated data collections available. To retrieve our training and test data we used DiAML querying language and procedures proposed in Petukhova et al. (2014). Figure 1 illustrates what type of data is extracted from each corpus. Dialogue acts of two types are considered: *information transfer* and *action discussion* acts. Information transfer acts are used to obtain (information-seeking) or to provide information (information providing). We extracted all statements and answers, and merged them into an **Inform** class. This is logically eligible operation since all information-providing acts have in common that the speaker provides the addressee certain information which he believes the addressee not to know or not to be aware of, and which he assumes to be correct. The various subtypes of information-providing acts differ in the speaker's motivation for providing the information, and in different additional beliefs about the information that the addressee possesses.

Information-seeking acts comprise all types of **Questions** including Propositional Question, Check Question, Set Question and Choice Question. All acts in this class have in common that the speaker wants to know something, which he assumes the addressee to know, and puts pressure on the addressee to provide this information.

Action discussion acts have a semantic content consisting of an action, and possibly also a description of a manner or frequency of performing the action. Request, Instruct, Suggestion and Accept/Reject Offer acts belong to a class of **Directives** and are concerned with the speaker's wish that the addressee performs an action. These acts are distinguished by the degree of pressure that the speaker puts on the addressee and the speaker's assumptions about the addressee's ability and agreement to perform a certain actions.

Commissive acts such as Accept/Reject Request or Suggestion and Offer capture the speaker's commitments to perform certain actions. Thus, in our data sets we have four main dialogue act categories as classes: Inform, Question, Commissives and Directives acts.

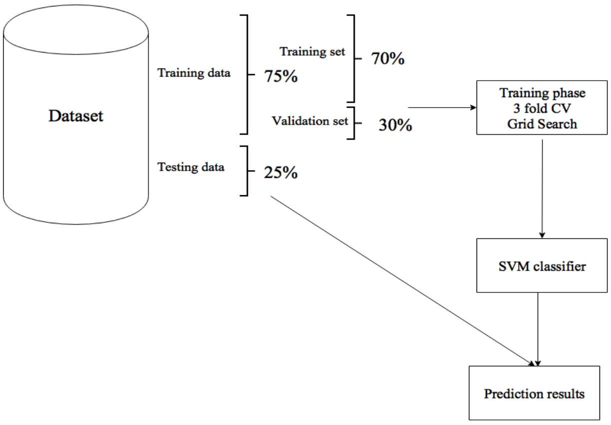


Figure 3: Classification process: training, development and test data partition.

The distinguished dialogue act types frequently occur in the analysed corpora (see Table 1 for tags distribution in all considered corpora). They frequently occur in our new collected dialogue data, the METALOGUE corpus[4], see (Petukhova et al., 2015) and (Petukhova et al., 2016). They are also expected to be frequent in any dialogue data, although with different class distributions. For the classifiers to learn dialogue act classes under realistic conditions, negative examples (**Other** class) were added to our data set. The total number of instances extracted from the AMI corpus are 19.918 for all classes; from the HCRC MapTask corpus 2.920; and from the Switchboard 12.768. METALOGUE corpus contains 1.520 dialogue act instances.

Since AMI, HCRC MapTask and Switchboard dialogues are segmented into units that are continuous stretches of

[3] For a detailed comparison of the schemes see (Petukhova, 2011).

[4] More information about the METALOGUE project can be found at www.metalogue.eu

		Training set						
		AMI	MapTask	SWBD	SWBD+MapTask	AMI+SWBD	AMI+MapTask	AMI+SWBD+MapTask
Test set	AMI	-	0.58	0.77	0.78	-	-	-
	MapTask	0.54	-	0.53	-	0.56	-	-
	SWBD	0.86	0.61	-	-	-	0.81	-
	Metalogue	0.84	0.66	0.84	0.84	0.83	0.84	0.83

Table 3: Classifiers performance in terms of F-scores on cross-corpora training and testing.

Features set	unigrams	bi-grams	tri-grams
Chunks	0.45	0.71	0.41
Chunks, POS	0.63	0.75	0.55
Chunks, word tokens	0.66	0.68	0.60
Chunks, POS, word tokens	0.79	**0.84**	0.74
POS	0.62	0.58	0.64
POS, word tokens	**0.82**	0.79	0.76
word tokens	0.74	**0.81**	0.67

Table 2: Dialogue act classification results in terms of F-score on different feature set and with n-gram range computed(best results reported).

communicative behaviour to which only one dialogue act tag is assigned, we cleaned the data by removing all segment internal hesitations and disfluencies. METALOGUE data was segmented in multiple dimensions using the approach reported in (Geertzen et al., 2007) resulting in minimal meaningful segments[5] that do not contain any irrelevant material such as, for example, turn internal stallings, restarts or other flaws in speech production.

4. Classification Experiments Set Up

In order to train classifiers that are able to operate on data collected in various domains, along with commonly used n-grams and bag-of-words models, we used Part-of-Speech (POS) information and shallow syntactic parsing features, and combinations of those. Linguistic features are expected to contribute to higher cross-domain portability of trained prediction models. For POS tagging the Stanford CoreNLP[6] tagger was used and chunking was performed using the Illinois shallow parser (Punyakanok and Roth, 2001).

Stratified cross-validation[7] classification experiments (see Figure 3) were carried out

1. using different feature sets to assess the features' importance for a given task;
2. varying train and test data set partitions to assess the flexibility of each prediction model to deal with data from a different domain; and
3. splitting up the output structure to predict individual class labels, semantically opposite classes (e.g. Question vs Inform) and all classes together having binary and multi-class training.

There were seven basic training experimental set ups defined. Additionally, all experiments were repeated exploiting corpora as training and testing sets disjointedly, pairwise, and together as one big training and testing set. Note that testing was performed on the corpus data that was not present in training set and therefore unseen by the classifier. Additionally, all trained models were tested on the METALOGUE data. Figure 2 illustrates our training/testing design.

5. Results: Cross-Corpora Training and Testing

Support Vector Machine (SVM) (Boser et al., 1992) classifier training was performed using the scikit-learn implementation[8]. In all experiments, radial basis function kernels has been used. The most commonly used performance metrics such as accuracy, precision, recall and F-scores (harmonic mean) were computed to evaluate the classifiers' performance.[9] It should also be noted here that when running cross-corpora training/testing experiments data has been resampled based on the distribution of test data target dialogue acts in the first set of experiments. Experiments were repeated using a weighted average for an unknown target domain.

5.1. Feature Engineering and Data Sets Effects

As for features, the best results were obtained on the complex features combining bigrams of POS tags, chunking information and word tokens (see Table 2). When trained on unigram word token models only we observed a decrease in performance of about 10% compared to the performance using the combined features; trained on n-gram POS tag models 5% on average; and trained on n-gram chunking information 20%. Thus, wording of an utterance is still very important, but, when supplied with linguistic information, the performance of the classifier improves.

It was also observed that the performance drops significantly when trained on the MapTask corpus (more than 40%), while linguistically rich corpora like AMI and Switchboard have stronger predictive power for the new METALOGUE corpus (see Table 3). This can be attributed to the representation of the utterances in the dataset (i.e.

[5] A functional segment in ISO 24617-2 is defined as a minimal stretch of communicative behaviour that has a communicative function (and possibly more than one).

[6] http://nlp.stanford.edu/software/corenlp.shtml

[7] Stratified cross-validation is used when classes distribution is unbalanced. It helps account for the randomness of the original training data sample. Each fold in our experiments has the same class distribution, since we were not sure that our original training set had a representative class distribution for the target data. Stratified cross-validation reduces the variance of the estimates and improves the estimation of the generalization performance of classifier algorithms.

[8] http://scikit-learn.org/stable/

[9] For the sake of simplicity in this paper we report F-scores. Other values are available upon request.

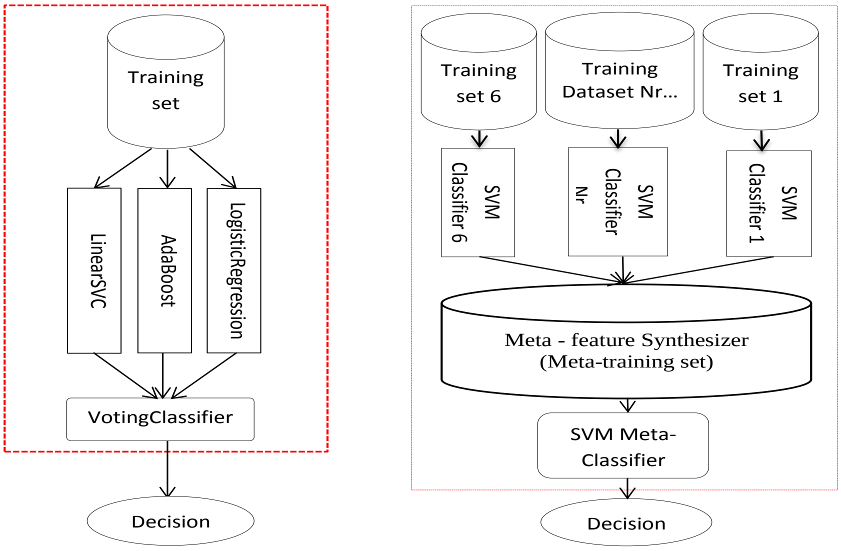


Figure 4: Majority Voting Classification (left) and SVM-based Meta-Classification (right) procedures.

the length of the utterances in the MapTask corpus). However, when the classifier is trained on the combination of the MapTask corpus with other corpora, the performance of the classifier improves. Therefore, we concluded that corpus data with a broad spectrum of dialogue phenomena like AMI and Switchboard conversations can be profitably used to train models that have stronger predictive power when applied to new dialogue data from other domains, while MapTask based models are rather limited in this respect.

5.2. Dialogue Act Models Adaptation

When new dialogue data is collected and further processing requires dialogue act annotation and analysis, we propose two main methods to obtain such annotations automatically (or semi-automatically) . The first one can be applied if a small annotated in-domain corpus is available and/or easy to obtain. The second approach can be used when no annotated data is accessible.

Following the first approach, the small amount of annotated data can be provided to the classifier as training/testing set, and cross-validation learning experiments can be performed. However, since the unknown data does not necessarily share the same feature space and the same distribution as the known, labelled data, the following should be taken into account:

1. the selection of the known out-of-domain data (source data) is not random but justified, e.g. if the properties of the new data (target data) are known, the semantically and pragmatically closest possible corpus should be selected based , for example, on semantic similarity, dialogue act tags distribution drawn on relative frequencies and/or dialogue act sequences comparative analysis;
2. the out-of-domain data is re-sampled, selecting instances based on the target in-domain-data distribution or the unknown target in-domain-data is re-weighted;
3. the feature representation is adjusted to reduce differences between source and target domains, e.g. the use of domain-independent features such as linguistic information rather than domain-dependent ones such as wording in our task.

Following the second approach, the dialogue act prediction models can be selected that show the best performance on other corpora or on the largest corpus. After applying them to the unseen data, output predictions can be used as pre-annotated data for manual correction performed by trained annotators. Post-correction of the pre-annotated data set will reduce overall annotation time. The correct annotated corpus data can then be used in the second training iteration as described in the first procedure. These steps can be repeated several times till the classifier achieves satisfactory performance, and/or any further improvements are impossible.

5.3. Late Fusion

As shown in the previous sections, given certain training conditions (different feature sets and training data partitions), we obtained possible output predictions (hypotheses) from our classifiers. As a follow-up step, the decision has to be taken what features and training set constellation gives rise to the best performance for a new collected dataset, taking into account various factors described in the previous subsection. To automatize and optimize this process, at the decision level the *late fusion* methods can be used when combining the prediction scores available from multiple classifiers. We tested two possible late fusion alternatives when deciding on the strongest prediction model for the new collected corpus: *majority voting* (Morvant et at, 2014) and/or *meta-classification* (Lin and Hauptmann, 2002). Figure 4 illustrates both decision taking set ups. In both cases multiple classifiers predictions are used as valid hypotheses. A *hard voter* simply counts hypotheses and decides on the one winning one. Additionally, classifiers confidence scores can be taken into account. For example, a *soft voter* takes predicted class probabilities into account,

Late fusion method	(Meta-)training sets						
	AMI	MapTask	SWBD	AMI+MapTask	AMI+SWBD	MapTask+SWBD	AMI+MapTask+SWBD
Majority Voting	0.76	0.59	0.77	0.77	0.73	0.78	0.72
Meta-classification	0.86	0.85	0.83	0.85	0.85	0.85	0.85

Table 4: Late fusion classifiers performance in terms of F-scores on different training sets with Metalogue corpus as a test set.

and the final class label is derived from the class label with the highest average probability. In our experiments we tested a hard voter. The majority voting method allows the incorporation of different types of classifiers, e.g. Logistic Regression (Yu et al, 2011), AdaBoost (Zhu et al, 2009) and the Linear Support Vector Classifier (LinearSVC, see Vapnik, 1995) as in our experiments.

Majority voting is a rather straightforward and robust method but may have some drawbacks, since some hypotheses can be wrong and, if in majority, they will lead to a wrong final decision. It has been shown in previous studies that the prediction history and current best prediction may be taken as features in the meta-classification step and help to discover and correct errors, see e.g. ‘recurrent sliding window strategy’ (Dietterich, 2002) or adaptive training (Van den Bosch, 1997). Meta-classification is a powerful approach to boost any classification system. All *local* classifiers are trained as described above. The final decision is, however, in the hands of the meta-classifier (also SVM-based) which, in addition to all local classifiers’ features, gets their prediction history as an input and re-classifies in new feature space.

The obtained results indicate that the meta-classifier has stronger and more stable prediction power than the majority voting classifier (Table 4). The obtained F-scores ranging between 0.83 and 0.86 are comparable to those of the best local classifiers. Even weak local MapTask models do not affect the overall meta-classification performance, see Table 3 for comparison. The conclusion was drawn here is that the classification procedure can be optimized by using a meta-classifier in the final decision taking step. This may reduce design, training and analysis costs significantly.

6. Conclusions and Future Work

In this paper we discussed procedures to obtain new annotated dialogue data by using available resources and supervised machine learning algorithms. We tested suggested procedures on different data sets and various features. We specified requirements on two prediction models adaptation settings. We also concluded that the meta-classifier combination strategy is superior to individual classifiers and outperforms the majority voting strategy. The meta-classification combination strategy improves classification accuracy, even when using limited and semantically different training data.

We believe that, applying the proposed methods, new semantically annotated resources can be created on a rather large scale. The approach is very promising and the resulting data is expected to be of sufficiently high quality compared to noisy data obtained applying unsupervised methods or methods that involve distant supervision. Another strong expectation is that annotation costs will be reduced by a minimum of 40% given the average F-scores of 0.85 obtained in our cross-domain/corpora classification experiments. We showed that the use of a meta-classifier eliminates the problem of model adaptation when re-sampling the training set and assigning weights to individual classes. In the future, we plan to test deep learning methods for dialogue act recognition, which require less feature engineering efforts (Henderson et al., 2014). To further automatize the annotation process, plug-ins for dialogue annotation tools, e.g. ANVIL[10] and ELAN[11], will be designed.

Acknowledgements

The reported research is partly funded by the EU FP7 Metalogue project, under grant agreement number: 611073. We are also very thankful to anonymous reviewers for their valuable comments.

7. Bibliographical References

Blache, P., Bertrand, R.m and Ferré, G. 2009. Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project. In Multimodal Corpora, Springer-Verlag.

Van den Bosch, A. 1997 Learning to pronounce written words: A study in inductive language learning. PhD thesis, Maastricht University, The Netherlands

Boser, B., Guyon, I., and Vapnik, V. 1992 A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*,Pittsburgh.

Bunt, H., Fang, A., Cao, J., Liu, X., and Petukhova, V. 2013 Issues in the addition of ISO standard annotations to the Switchboard corpus. In *Proceedings 9th Joint ISO-ACLWorkshop on Interoperable Semantic Annotation (ISA-9)*, Potsdam.

Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. 1996 *HCRC Dialogue Structure Coding Manual.* Human Communication Research Centre HCRC TR-82, University of Edinburgh.

Di Eugenio, B. et al. 1998 *The COCONUT project: dialogue annotation manual.* ISP Technical Report 98-1.

Dielmann, A. and Renals, S. 2008 Recognition of Dialogue Acts in Multiparty Meetings Using a Switching DBN. In Audio, Speech, and Language Processing, IEEE Transactions, pp. 1303-1314

Dietterich, T. 2002 Machine learning for sequential data: a review. Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, pp. 15-30

Dhillon, R. et al. 2004 Meeting recorder project: dialogue labelling guide. ICSI Technical Report TR-04-002

[10] `http://www.anvil-software.org/`

[11] `https://tla.mpi.nl/tools/tla-tools/elan/`

Fang, A., Cao, J., Bunt, H., and Liu, X. 2012 The annotation of the Switchboard corpus with the new ISO standard for Dialogue Act Analysis. In *Proceedings 8th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-8)*, Pisa.

Geertzen, J., Petukhova, V., and Bunt, H. 2007. A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification. *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, pages 140–149.

Henderson, M., Thomson, B., and Young, S. 2014. Word-Based Dialog State Tracking with Recurrent Neural Networks. *Proceedings of the SIGdial*, Philadelphia.

Jurafsky, D., Schriberg, E., and Biasca, D. 1997 *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation: Coders Manual*, University of Colorado.

Jurafsky, D. et al. 1998 Lexical, prosodic, and syntactic cues for dialogue acts In Proceedings of the Discourse Relations and Discourse Markers Conference, Somerset, New Jersey, USA

Fernandez, R. and Picard, R. 2002 Dialog act classification from prosodic features using Support Vector Machines In Proceedings of Speech Prosody 2002, Aix-en-Provence, France

ISO. 2012 *Language resource management – Semantic annotation framework – Part 2: Dialogue acts.* ISO 24617-2, Geneva, ISO, Geneva.

Keizer, S. 2003 Reasoning under uncertainty in natural language dialogue using Bayesian Networks. PhD Thesis,Twente University Press,The Netherlands

Lendvai, P. et al. 2004 Memory-based Robust Interpretation of Recognised Speech. In Proceedings of the SPECOM '04, St. Petersburgh, Russia, pp. 415-422

Lin, W.-H., and Hauptmann, A. 2002 Meta-classification: Combining multimodal classifiers. Mining Multimedia and Complex Data. Springer Berlin Heidelberg, pp. 217-231.

Morvant, E., Habrard, A., and Ayache, S. 2014 Majority vote of diverse classifiers for late fusion. Structural, Syntactic, and Statistical Pattern Recognition. Springer Berlin Heidelberg, pp. 153-162.

Petukhova, V. and Bunt, H. 2007 A multidimensional approach to multimodal dialogue act annotation. In Proceedings of the IWCS 2007, pp. 142-153

Petukhova, V. 2011 Multidimensional Dialogue Modelling. PhD dissertation. Tilburg University, The Netherlands

Petukhova, V., Malchanau, A., and Bunt, H. 2014 Interoperability of dialogue corpora through ISO 24617-2-based querying. In Proceedings of the LREC 2014, Iceland.

Petukhova, V., et al. 2014 The DBOX corpus collection of spoken human-human and human-machine dialogues. In Proceedings of the LREC 2014, Iceland.

Petukhova, V., et al. 2015 Modelling argumentation in parliamentary debates. In Proceedings of the 15th Workshop on Computational Models of Natural Argument, Principles and Practice of Multi-Agent Systmes Conference (PRIMA 2015), Bertinoro, Italy

Petukhova, V., et al. 2016 Modelling multi-issue bargaining dialogues: data collection, annotation design and corpus. In Proceedings of the LREC 2016, Portorož, Slovenia

Popescu-Belis, A. 2003 Dialogue act tagsets for meeting understanding: an abstraction based on the DAMSL, Switchboard and ICSI-MR tagsets. Project Report IM2.MDM-09, v1.2, December 2004

Punyakanok, V., and Roth, D. 2001 The Use of Classifiers in Sequential Inference. NIPS, pp.995–1001

Reithinger, N.and Klesen, M. 1997 Dialogue act classification using language models. In Proceedings of EuroSpeech-97, pp. 2235-2238

Samuel, K. et al. 1998 Dialogue act tagging with transformation-based learning. In Proceedings of the ACL and CoLing, Montreal, pp. 1150 - 1156

Shriberg, E. et al. 1998 Can prosody aid the automatic classification of dialog acts in conversational speech? In Language and Speech (Special Issue on Prosody and Conversation), 41(3-4):439-487

Stolcke, A. et al. 2000 Dialogue act modeling for automatic tagging and recognition of conversational speech. In Computational Linguistics, 26(3), pp. 339-373

Surendran, D., and Levow, G. 2006 Dialog act tagging with Support Vector Machines and Hidden Markov Models. In Proceedings of Interspeech/ICSLP

Vapnik, V.N. 1995 The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.

Webb, N. et al. 2005 Error Analysis of Dialogue Act Classification. In Proceedings of the TSD Conference, Karlovy Vary, Czech Republic, pp. 451-458

Webb, N., and Liu, T. 2008 Investigating the Portability of Corpus-derived Cue Phrases for Dialogue Act Classification. In Proceedings of the COLING '08, Manchester, United Kingdom

Webb, N., and Ferguson, M. 2010 Automatic Extraction of Cue Phrases for Cross-Corpus Dialogue Act Classification. In Proceedings of the COLING '10, pp. 1310-1317

Yu, H.-F., Huang, F.-L., Lin, C.-J. 2011 Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning 85(1-2):41-75.

Zhu, J., Zou, H., Rosset, S., and Hastie, T. 2009 Multi-class AdaBoost.

8. Language Resource References

AMI-Consortium. 2005 AMI Meeting Corpus. available at `http://groups.inf.ed.ac.uk/ami/download/`

Anderson, A., et al. 1991 The HCRC Map Task Corpus. Language and Speech, 34, pp. 351-366. Available at `http://groups.inf.ed.ac.uk/maptask/`

John, G., and Holliman, E. 1993 Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium.