Transfer of Corpus-Specific Dialogue Act Annotation to ISO Standard: Is it worth it?

Shammur Absar Chowdhury, Evgeny A. Stepanov, Giuseppe Riccardi
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Italy
{sachowdhury, stepanov, riccardi}@disi.unitn.it

Abstract

Spoken conversation corpora often adapt existing Dialogue Act (DA) annotation specifications, such as DAMSL, DIT++, etc., to task specific needs, yielding incompatible annotations; thus, limiting corpora re-usability. Recently accepted ISO standard for DA annotation – Dialogue Act Markup Language (DiAML) – is designed as domain and application independent. Moreover, the clear separation of dialogue dimensions and communicative functions, coupled with the hierarchical organization of the latter, allows for classification at different levels of granularity. However, re-annotating existing corpora with the new scheme might require significant effort. In this paper we test the utility of the ISO standard through comparative evaluation of the corpus-specific *legacy* and the semi-automatically transferred *DiAML* DA annotations on supervised dialogue act classification task. To test the domain independence of the resulting annotations, we perform cross-domain and data aggregation evaluation. Compared to the *legacy* annotation scheme, on the Italian LUNA Human-Human corpus, the DiAML annotation scheme exhibits better cross-domain and data aggregation classification performance, while maintaining comparable in-domain performance.

Keywords: Dialogue Act, Dialogue corpora, Speech, DiAML, ISO Standard

1. Introduction

Dialogue Acts (DA) are fundamental for the analysis of conversations: they carry communicative functions such as question, answer, expression of agreement and disagreement, etc. Consequently, the range of applications of DA analysis is quite wide and includes conversation summarization (both spoken and written), dialogue systems, etc.; and DAs have been extensively studied in both theoretical and computational linguistics. In the absence of a single commonly accepted standard, spoken corpora often adapt existing domain independent annotation schemes like DAMSL (Core and Allen, 1997), TRAINS (Traum, 1996), DIT++ (Bunt, 2005) to task-specific needs; thus, creating incompatible annotations. The supervised and unsupervised annotation and classification of DAs (e.g. (Joty et al., 2011)) and cross-domain and cross-media classification (e.g. forums, email, and spoken conversations (Joty et al., 2011; Tavafi et al., 2013)) have single important drawback: since the sets of considered DAs are not consistent, introduced cross-corpora mappings are at best generalizations or subsets.

Recently accepted international ISO standard for DA annotation – Dialogue Act Markup Language (DiAML) (Bunt et al., 2010; Bunt et al., 2012) – could serve as a *lingua franca* for cross-corpora DA mapping. However, such mappings might require significant amount of manual re-annotation effort. The utility of abandoning the *legacy* annotations and manually or semi-automatically re-mapping them to the ISO standard could be tested under two conditions: (1) if the new annotation is equal or superior in supervised DA classification and (2) if it is indeed domain independent and allows both cross-domain application and pooling data from different domains (i.e., data aggregation). Thus, in this paper we presents experiments on the semi-automatic re-annotation of the Italian LUNA Corpus (Dinarelli et al., 2009) with DiAML and evaluation of the annotations in cross-domain and data aggregation settings.

Dimension	ABBR	ISO	LUNA
General (Task)	G	26	8
Social Obligations Management	SOM	10	4
Auto-Feedback	AutoFb	2	
Allo-Feedback	AlloFb	3	3
Time Management	TimeM	2	
Turn Management	TurnM	6	–
Discourse Structuring	Disc	2	–
Own Speech Management	OSM	2	–
Partner Speech Management	PSM	3	–
Total		56	15

Table 1: Mapping LUNA dialogue acts to DiAML ISO Standard 9 dialogue act dimensions and communicative functions with counts per dimension.

In the rest of the paper we describe the LUNA to ISO DA Mapping (Section 2.) and the annotation procedure (Section 3.). In Section 4. we report on the supervised DA classification experiments comparing the *legacy* and ISO annotation schemes; and the cross-domain performance of the new annotations. Section 5. provides concluding remarks.

2. Mapping LUNA to ISO Standard

The Italian LUNA Human-Human Corpus (Dinarelli et al., 2009) is a collection 572 dialogues in the hardware/software help desk domain. The dialogues are conversations of the users engaged in problem solving task. A subset of 50 dialogues was annotated with dialogue acts. The LUNA DA annotation scheme (Quarteroni et al., 2008) was inspired by DAMSL (Core and Allen, 1997), TRAINS (Traum, 1996), and DIT++ (Bunt, 2005). The most common 15 dialog acts from these taxonomies are grouped into three categories (Dinarelli et al., 2009): *Core Dialog Acts* (8) are main actions in the dialog, such as request of information, response, or performing the task; *Conven-*

tional/Discourse Management Acts (4) are utterances such as greetings, apologies, etc. whose function is to maintain general dialog cohesion; *Feedback/Grounding Acts* (3) are utterances whose function is to acknowledge, provide feedback, or just time fillers; and *Others* (1) to capture the rest. The unit of annotation for dialogue acts in LUNA Corpus is an utterance. However, due to the overlapping turns (both speakers speaking), an utterance can span several turns. Thus, the dialogue act annotation was preceded by additional utterance segmentation.

Full description of the DiAML annotation scheme (Bunt et al., 2012) is out of the scope of this paper. Rather we focus on the DA tag set and dimensions. The DiAML annotation scheme consists of 56 DA tags (communicative functions), organized into 9 dimensions: 26 general (applicable to any dimension) and 30 dimension specific (see Table 1, ISO column).

The issues of converting DAMSL-based corpus to the ISO standard were addressed by (Fang et al., 2012) and (Bunt et al., 2013). Following the re-annotation methodology outlined in (Fang et al., 2012) we mapped LUNA DAs to DiAML. LUNA contains only 15 tags compared to DiAML's 56, and most of the relations in the mapping are one-to-many. Even though, some of these relations can be disambiguated with respect to context (Petukhova et al., 2014) (e.g. if the DA in the previous turn is *Info-Request* and the current DA is *Yes-Answer*, there is a high chance that the former maps to *Propositional Question* and the latter to *Confirm*), since both relations are one-to-many, such mapping is error prone. Thus, automatic mapping is manually examined. Due to data distribution and for the consistency with the legacy annotation, we did not annotate all the dimensions: Discourse Structuring, Speech and Turn Management dimensions were mapped to *Other*.

For cross-domain experiments, on the other hand, a set of 10 call center dialogues was sampled from large scale call centers conversations providing customer care support in energy and utilities domain. This set is annotated with DiAML scheme only.

3. Re-Annotation Methodology

In (Bunt et al., 2013) the authors list segmentation differences as one of the issues of converting DAMSL-based annotation to ISO standard. While in the former the unit of annotation usually corresponds to a turn, in the latter it is a *functional segment* that can be shorter or longer than the turn. In LUNA, on the other hand, the unit of annotation was considered to be an utterance, which is similar to turn, ignoring the other speaker barge-ins. Consequently, re-annotation procedure also included re-segmentation.

As the first step of the re-annotation effort, a linguist annotated a limited set of LUNA dialogues to get accustomed to the procedure. Since the legacy annotation was performed by a different person, to ensure the consistency, the annotator performed an **unsupervised** annotation (15 dialogues) of the LUNA corpus with new DiAML scheme in the dimensions selected previously. This set of 15 dialogues is used to compute the inter-annotator agreement between the 'legacy' and the 'ISO' annotator.

For the agreement calculation ISO DAs are mapped to the 'legacy' DAs. Due to segmentation differences the two annotations are first aligned with respect to the Levenshtein distance and F-measure is computed with respect to alignment errors (Chowdhury et al., 2015). Since 'legacy' annotation unit covers several functional segments, insertion errors are ignored. The overall agreement between the 'legacy' and ISO annotators is $F_1 = 0.68$.

As the second step, we have annotated 10 dialogues from an out-of-domain corpus. The activity has two goals: (1) to check the dimension and DA distributions cross-domain and (2) for later cross-domain evaluation on supervised classification task. The resulting annotation was compared to the random 10 dialogues from LUNA annotation, from the previous step. The dimension and communicative function distributions were observed to be similar.

As the third step, the remaining LUNA dialogues are automatically re-annotated using the mapping described in Section 2., which was refined through steps 1 and 2. The annotator's job at this step was to segments the turns into functional units and to disambiguate the labels. This step is a **supervised** annotation, and automatic mapping is provided to ensure consistency with the **unsupervised** annotation, while reducing the amount of the required effort.

The distribution of the resulting annotation into dimensions is given in Table 3 together with the train and test splits. In the next section, we evaluate the annotation on the supervised DA classification task using this split.

4. Supervised Classification Experiments and Results

For the dialogue act classification, we use Sequential Minimal Optimisation (SMO), a support vector machine implementation, with its linear kernel and default parameters (Hall et al., 2009). As it was already mentioned, the 'legacy' and 'ISO' annotations are evaluated in three settings: (1) in-domain, (2) cross-domain and (3) data aggregation. We perform classification into dimensions and into communicative functions, using bag-of-words representation for features. The distribution of labels in each layer (dimensions and communicative functions) is unbalanced (see Table 3); however, we do not address balancing issues. For consistency with the 'legacy' annotation, we merged *Feedback* and *Time Management* dimensions. The *Social Obligations Management* dimension was kept separate. Performance is evaluated using standard precision, recall and F_1.

The results of the experiments on the dialogue act classification at dimension level are reported in Table 4 as F_1. In dimension level classification, the number of classes (dimensions) is the same for the 'legacy' and ISO annotated data. Due to the segmentation differences, the number of instances, however, is different. The results illustrate that in-domain performances of the two annotation schemes are comparable; however, ISO annotation scheme has better performances in the cross-domain and data-aggregation settings.

Communication function level classification settings are different for the 'legacy' and ISO annotated data: for the former it is classification into 16 classes, and for the latter

LUNA DA	ISO DA
Core Dialogue Acts → General/Task (**G**)	
Info-Request	Question, Set-Question, Choice-Question, Propositional-Question, Check-Question
Action-Request	Instruct, Suggest, Request
Yes-Answer	Confirm, Accept-Offer, Accept-Request, Accept-Suggest
No-Answer	Disconfirm, Decline-Offer, Decline-Suggest, Decline-Request
Answer	Address-Offer, Address-Request, Address-Suggest, Answer, Correction, Disagreement, Agreement
Offer	Offer, Promise
Report-On-Action	Inform
Inform	Inform, SOM:I-Self-Introduction, SOM:R-Self-Introduction
Conventional Dialogue Acts → Social Obligations Management (**SOM**)	
Greet	I-Greeting, R-Greeting
Quit	I-Goodbye, R-Goodbye
Apology	Apology, Accept-Apology
Thank	Thanking, Accept-Thanking
Feedback/Turn Management Dialogue Acts	
Clarif-Request	AlloFb:Positive, AlloFb:Negative
Ack	AutoFb:Positive, AutoFb:Negative
Filler	TimeM:Stalling, TimeM:Pausing
Non-Interpretable/Non-Classifiable Dialogue Acts	
Other	Other

Table 2: Mapping from LUNA DA to ISO dimensions and communicative functions. Note that most of the relations are one-to-many and frequently are cross-dimension.

Dimension	**Train** (40)	**Test** (10)	**Total** (50)	**OOD** (10)
General (Task)	1,456 (74.7%)	494 (25.3%)	1,950 (59.7%)	911 (61.9%)
Social	197 (78.8%)	53 (21.2%)	250 (7.6%)	99 (6.7%)
Auto-Feedback	530 (78.8%)	143 (21.2%)	673 (20.6%)	278 (18.9%)
Allo-Feedback	36 (81.8%)	8 (18.2%)	44 (1.3%)	11 (0.75%)
Time Management	74 (64.9%)	40 (35.1%)	114 (3.5%)	68 (4.62%)
Other	154 (65.0%)	83 (35.0%)	237 (7.3%)	105 (7.13%)
Total	2,447 (74.9%)	821 (25.1%)	3,268 (100.0%)	1,472 (100.0%)

Table 3: Distribution of dialogue acts in LUNA corpus and the out-of-domain corpus (**OOD**). The counts are given per annotated dimension and in total.

	Legacy			**ISO**		
Dimension	**ID**	**XD**	**AG**	**ID**	**XD**	**AG**
Task	0.79	0.72	0.80	0.78	0.74	0.80
Social	0.86	0.66	0.78	0.84	0.78	0.83
Time+Fb	0.71	0.61	0.69	0.73	0.64	0.72
Other	0.18	0.15	0.15	0.24	0.22	0.26
Micro	0.72	0.60	0.72	0.72	0.67	0.74

Table 4: Comparative evaluation of 'legacy' and ISO annotations at dimension level. F_1 for in-domain (**ID**), cross-domain (**XD**), and data-aggregation (**AG**) evaluation settings.

into 41 class. To evaluate the ISO scheme in more comparable settings, we additionally evaluate it after mapping to the 'legacy' annotation (i.e., to 16 'legacy' classes). The results of the experiments on the dialogue act classification at communication function level are reported in Table 5 as F_1. Individual communication function performances are aggregated to dimension level and reported numbers are micro-average F_1s. Comparing cross-domain and data aggregation setting results for ISO and ISO mapped to 'legacy' settings, we can observe that the models trained on ISO annotated data perform better on out-of-domain data. Increasing the training data size with the out-of-domain data also appears to be beneficial.

The 'legacy' annotation trained models benefit from the aggregation of the out-of-domain data to a greater extent than the ISO models. However, due to the greater drop in cross-domain evaluation of the 'legacy' models and better in-domain performance of the mapped ISO annotation, we conclude that the transfer of legacy annotation to the ISO standard is beneficial.

5. Conclusion

In this paper we have presented the semi-automatic process of transferring corpus-specific dialogue act annotation to the recently accepted ISO standard. The utility of the effort is assessed by evaluation of the 'legacy' and ISO annotations in in-domain, cross-domain and data aggregation settings. We have observed that the ISO annotation scheme exhibits better cross-domain and data aggregation performance for the task of supervised dialogue act classification at dimension level. For the communicative function level classification, we also have observed that the new annotation scheme provides better cross-domain generalization. Thus, indeed, it is worth transferring legacy annotations to the ISO standard.

The supervised dialogue act classification experiments have only utilized dialogue act span tokens in the bag-of-words

	Legacy (16)			**ISO** (41)			**ISO Mapped** (16)		
Dimension	**ID**	**XD**	**AG**	**ID**	**XD**	**AG**	**ID**	**XD**	**AG**
Task	0.31	0.20	0.38	0.24	0.27	0.26	0.35	0.39	0.36
Social	0.64	0.39	0.73	0.60	0.41	0.62	0.70	0.52	0.78
Time+Fb	0.68	0.55	0.68	0.84	0.63	0.83	0.84	0.62	0.83
Other	0.25	0.26	0.35	0.22	0.26	0.25	0.24	0.36	0.27
Micro	0.44	0.30	0.49	0.40	0.37	0.41	0.47	0.45	0.48

Table 5: Comparative evaluation of 'legacy' and ISO annotations at communicative function level. F_1 for in-domain (**ID**), cross-domain (**XD**), and data-aggregation (**AG**) evaluation settings. For each annotation scheme, the number of communicative functions is reported in parentheses. **ISO Mapped** reports performance of the ISO annotations after mapping to the 'legacy' annotation.

setting, which we consider to be a baseline. Since communicative functions are contextual, in the future we plan to experiment with the context for the classification.

6. Acknowledgements

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007–2013) under grant agreement 610916: SENSEI.

7. Bibliographical References

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., et al. (2010). Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. R. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Bunt, H., Fang, A. C., Liu, X., Cao, J., and Petukhova, V. (2013). Issues in the addition of ISO standard annotations to the switchboard corpus. In *Workshop on Interoperable Semantic Annotation*.

Bunt, H. (2005). A framework for dialogue act specification. In *In Proceedings of SIGSEM WG on Representation of Multimodal Semantic Information*.

Chowdhury, S. A., Calvo, M., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., García, F., and Sanchis, E. (2015). Selection and aggregation techniques for crowdsourced semantic annotation task. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Core, M. G. and Allen, J. F. (1997). Coding dialogs with the damsl annotation scheme. In *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., and Riccardi, G. (2009). Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece.

Fang, A. C., Cao, J., Bunt, H., and Liu, X. (2012). The annotation of the Switchboard Corpus with the new ISO standard for dialogue act analysis. In *Workshop on Interoperable Semantic Annotation*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Joty, S., Carenini, G., and Lin, C.-Y. (2011). Unsupervised modeling of dialog acts in asynchronous conversations. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*.

Petukhova, V., Malchanau, A., and Bunt, H. (2014). Interoperability of dialogue corpora through ISO 24617-2-based querying. In *LREC*.

Quarteroni, S., Riccardi, G., Varges, S., and Bisazza, A. (2008). An open-domain dialog act taxonomy. Technical Report DISI-08-032, University of Trento, August.

Tavafi, M., Mehdad, Y., Joty, S., Carenini, G., and Ng, R. (2013). Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 117–121.

Traum, D. (1996). Conversational agency: The trains-93 dialogue manager. In *Proceedings of Twente Workshop on Language Technology, TWLT-II*.

8. Language Resource References

Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., and Riccardi, G. (2009). Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*.