ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents

Stefano Mezza, Alessandra Cervone, Giuliano Tortoreto,
Evgeny A. Stepanov, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy
name.surname@unitn.it

Abstract

Dialogue Act (DA) tagging is crucial for spoken language understanding systems, as it provides a general representation of speakers' intents, not bound to a particular dialogue system. Unfortunately, publicly available data sets with DA annotation are all based on different annotation schemes and thus incompatible with each other. Moreover, their schemes often do not cover all aspects necessary for open-domain human-machine interaction. In this paper, we propose a methodology to map several publicly available corpora to a subset of the ISO standard, in order to create a large task-independent training corpus for DA classification. We show the feasibility of using this corpus to train a domain-independent DA tagger testing it on out-of-domain conversational data, and argue the importance of training on multiple corpora to achieve robustness across different DA categories.

1 Introduction

The correct interpretation of the intents behind a speaker's utterances plays an essential role in determining the success of a dialogue. Hence, the module responsible for intents classification lies at the very core of many dialogue systems, both in research and industry (e.g. Alexa, Siri). Moreover, although the task of intent recognition is traditionally linked to task-based systems, recently it has also proved crucial for non task-based conversational systems. According to the results of the Amazon Alexa Prize challenge (Ram et al., 2017), the most successful conversational systems in the competition relied on a strong spoken language understanding module, while more than 60% of the approaches explicitly used intents.

Nevertheless, automatic intent recognition is hard, since participants' intents in a dialogue are implicit. Intent classification has therefore been mostly modeled as a supervised machine learning problem (Gupta et al., 2006; Xu and Sarikaya, 2013; Yang et al., 2017), with the consequent definition of intents taxonomies. Over time this led to the creation of expensive annotated resources (Price, 1990; Henderson et al., 2014) with the related time-consuming design of multiple intent schemes. In most cases, however, intents taxonomies are defined specifically for a given application or a dataset and are not generalizable to other systems or tasks, making these resources difficult to reuse (e.g. the popular Air Travel Information Services (ATIS) corpus include heavily domain-dependent intents such as *Airfare* or *Ground Service*).

Dialogue Acts (DA), also known as speech or communicative acts, represent an attempt to create a formalized and generalized version of intents. As such, DAs have been investigated by the research community for many years (Stolcke et al., 2000) and have been applied successfully to many tasks. In particular, their aspiration to generality makes them an appealing option for non task-based application (e.g. more than 20% of the teams in Amazon Alexa Prize Challenge explicitly used DAs (Cervone et al., 2017; Bowden et al., 2017), including the winning team (Fang et al., 2017)). Also, in the case of DAs, over the years there have been several efforts to produce publicly available annotated resources (Godfrey et al., 1992; Carletta, 2006; Alexandersson et al., 1998) to train DA taggers. The DA taxonomies created

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3539–3551
Santa Fe, New Mexico, USA, August 20-26, 2018.

for these resources, albeit arguably more general compared to corpora like ATIS (for example utilizing categories such as *wh-questions*), are still dataset specific; since many of these schemes lack coverage of some crucial aspects of dialogic interaction. Furthermore, given that all these datasets utilize different schemes, these resources are hardly compatible.

The ISO 24617-2 (Bunt et al., 2010; Bunt et al., 2012), the international ISO standard for DA annotation, represents the first attempt to create a truly domain and task independent scheme. Given its holistic approach compared to previous schemes, ISO 24617-2 can be used as a lingua franca for cross-corpora DA mapping, as confirmed by successful attempts to remap single corpora to the standard (Chowdhury et al., 2016; Fang et al., 2012).

However, there is no reference training set for the standard, since the only public resource currently available with ISO 24617-2 annotation (DialogBank, (Bunt et al., 2016)) is too small to be used to train classifiers. Therefore, most DA tagging research still focuses on in-domain studies on large datasets with incompatible DA annotations (Stolcke et al., 2000; Ji et al., 2016). Moreover, most publicly available corpora are imbalanced with respect to the distribution of various DA dimensions such as *Information Transfer* (e.g. "What's your favourite book?") or *Action Discussion* (e.g. "Tell me the news."), which are required for successful open-domain conversational systems.

In this work, we show how to reuse and combine publicly available annotated resources to create a large training corpus for domain-independent DA tagging experiments. We map different corpora using an ISO standard compliant DA taxonomy, following the previous research on the topic (Fang et al., 2012; Petukhova et al., 2014b) and we share this resource with the research community.[1]

In order to investigate the soundness of our approach compared to in-domain models we further experiment with domain-independent DA tagging. As previously done in the literature we cast the Dialogue Act tagging task as a supervised multi-class classification problem using Support Vector Machines. The correctness of the approach is first tested on the de facto DA tagging standard – the Switchboard (SWDA) corpus (Godfrey et al., 1992), using the reference training and test sets and achieving performance comparable to the state-of-the-art approaches. Secondly, we experiment with domain-independent DA tagging following the same approach and using our combined resource as a training corpus. The DialogBank corpus, that represents a reference manual DA annotation for the ISO standard, is used for the evaluation of the tagger. To the best of our knowledge this is the first attempt to test automatic DA annotation on this corpus.

The domain-independence and suitability of the tagger for conversational systems trained on multiple resources is additionally evaluated on two other corpora annotated following our optimized taxonomy (human-machine conversations from the Amazon Alexa Prize Challenge). The performances achieved on these three datasets suggest that the training on multiple corpora represents a step forward for DA tagging of open-domain non task-based human-machine conversations. Finally, we present experiments to investigate the contribution of the different corpora to the performance of the classifiers. The results of our experiments show the importance of utilizing multiple resources to achieve a sound performance across different types of DA categories. The multi-domain DA tagger presented here was successfully employed in Roving Mind, our open-domain conversational system for the Alexa Prize (Cervone et al., 2017).

2 State of the Art

2.1 Dialogue Act Annotation Schemes

The notion of Dialogue Acts can be traced back to the one of illocutionary acts introduced by (Austin and Urmson, 1962). The illocutionary act represents a level of description of an utterance's meaning that goes beyond the purely semantic level ("Is the window open?") to encompass the intent of the speaker in producing that utterance ("Please, close the window.").

One of the first DA taxonomies was the one created for the task-based corpus MapTask (Anderson et al., 1991) in the early nineties. The MapTask scheme distinguishes between *initiating moves* – such

[1] The suite of scripts we wrote to map and combine publicly available corpora can be found at `https://github.com/ColingPaper2018/DialogueAct-Tagger`

as giving instructions, explaining, checking information or asking questions – and *response moves* – for example acknowledging instructions, answering questions and clarifying information. The corpus also makes a distinction regarding the grammatical and semantic structure of the interactions, classifying, for example *wh-questions*, *yn-questions* and *positive/negative answers*. Although pioneering at the time, the MapTask annotation scheme is very specific to the described scenario, and some of its DAs (e.g. *instruct*, *clarify*, *check*) do not scale well to generic, non task-based conversations. Moreover, its taxonomy was not designed to capture all human behaviours during conversations, and, as a consequence, its coverage for labelling a non task-based interaction is inadequate.

The first attempt to define a unified, non task-based standard for DA tagging was the Discourse Annotation and Markup System of Labeling (DAMSL) (Core and Allen, 1997) tag-set for the SWDA (Godfrey et al., 1992) corpus. This annotation scheme proposes a taxonomy of 42 tags, describing both semantic aspects of conversation (*opinion*, *non-opinion*, *preference*, etc.), syntactic aspects (*yn-questions*, *wh-questions*, *declarative questions*, etc.) and behaviours related to the dialogue (*conventional closing*, *hedge*, *backchanneling*, etc.). Nevertheless, the taxonomy still has some issues: tags are mutually exclusive (making it impossible to annotate, for example, a no answer which was also signaling non-understanding) and are organised in a flat taxonomy, which does not take into account similarities and differences between the tags.

Bunt (1999) introduced the Dynamic Interpretation Theory (DIT) for dialogues, setting the theoretical foundation for a domain-independent and task-independent DA taxonomy. The paper introduced some very important concepts like the idea of *multidimensionality* of DAs and the distinction between *Action-Discussion*, a macro-category of DAs encompassing cases in which interlocutors negotiate actions to be performed (e.g. requests like “Let’s switch topic.”), and *Information-Transfer* interactions, capturing the DAs through which speakers exchange information (e.g. sharing personal information like “My name is John.”). The DIT++ taxonomy (Bunt, 2009) was then defined in 2009 with the aim of providing a unique and universally recognized standard for DA annotation based on the theoretical ideas introduced in the DIT scheme. Its fifth version was accepted as ISO 24617-2.

The core aspects of the ISO standard are its multidimensionality and its domain and task independence. The ISO scheme is multidimensional in the sense that it makes a clear distinction between *semantic dimensions* (i.e. the aspect of the communication which the DA describes) and *communicative functions* (i.e. the illocutionary act performed within that dimension). In this way, ambiguities between various aspects of the communication and overlapping between DAs are removed. Furthermore, the scheme contains a generic dimension and communicative functions, which is suitable for mapping virtually any kind of conversation, both task-based and non task-based. Moreover, its multidimensional aspect and hierarchical taxonomy make it extensible and potentially adaptable to specific conversational sets.

2.2 Dialogue Act Tagging

The automatic recognition of Dialogue Acts has been addressed by the literature using various machine learning techniques. In particular DA classification has been modeled both as a sequence labeling problem, using techniques such as HMM (Stolcke et al., 2000), neural networks (Ji et al., 2016; Kumar et al., 2017) or CRF (Quarteroni et al., 2011), and as a multi-class classification problem, using for example SVM (Quarteroni and Riccardi, 2010). Mentioning and comparing all DA classification approaches is difficult because of the differences in annotation schemes and datasets used. All approaches, however, are usually tested on in-domain data.

One of the most popular datasets for benchmarking is SWDA (Godfrey et al., 1992), a dataset of human-human open-domain telephone conversations. The state-of-the-art on SWDA (77.0% accuracy on 42 DA tags) was achieved by (Ji et al., 2016) using deep neural networks.

The ISO standard (Bunt et al., 2010) can be seen as a generalization of all these annotation schemes. However, there is no available training data for the ISO standard.

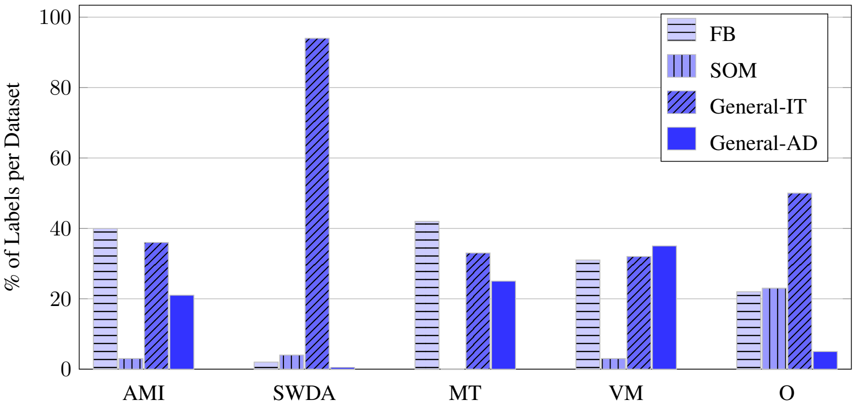

Figure 1: The distribution of dialogue act categories (after mapping to ISO standard) in various corpora – AMI, SWDA, MapTask (MT), VerbMobil (VM) and BT Oasis (O). The represented DA categories are Social Obligations Management (SOM) and Feedback dimension DAs, as well as *Action-Discussion* (AD) and *Information-Transfer* (IT) DAs from general dimension.

3 Data Sets

3.1 Training sets

The scarcity of resources of adequate size annotated with the ISO standard makes it difficult to train a DA tagger for this taxonomy. To the best of our knowledge, the DBOX corpus (Petukhova et al., 2014a) – the only resource manually annotated using the ISO standard – is not yet publicly available . The best possible approach given the current availability of data is to map existing corpora's DA schemes to the ISO scheme. Given the limited – and often domain-dependent – annotation scheme of these resources, it is impossible to map enough data to train a DA tagger for the full ISO taxonomy, since some of the ISO dialogue acts have no correspondence in any of the considered corpora. Therefore, we opted for a reduced version of the taxonomy, limiting our research to subsets of the General (Task), Social Obligation Management and Feedback dimensions. So far, we mapped the following five different corpora to our scheme:

SWDA: The Switchboard corpus (Godfrey et al., 1992) is a dataset of transcribed open-domain telephone conversations. The Switchboard Dialogue Act Corpus (SWDA) is a subset of the Switchboard corpus annotated with DAs. SWDA represents a logical choice when building a training set for a domain-independent DA tagger, as it is a large collection of open-domain, non task-based conversations, and therefore provides a natural similarity to the conversational domain of social bots. Moreover, there are already examples in literature of mappings from the Switchboard corpus to the ISO standard (Fang et al., 2012). As visible from Figure 3.1, drawbacks of the corpus with respect to the task include its unbalancedness (60% of utterances are *Information-providing*) and lack of *Action-Discussion* interactions (less than 1% of overall corpus).

AMI: This corpus contains transcriptions from 100 hours of meeting recordings of the European-funded AMI project (FP6-506811), a consortium dedicated to the research and development of technology (Carletta, 2006). This dataset presents a reasonably balanced collection of utterances and a taxonomy which shares some similarities with the ISO standard (e.g. distinction between *Action-discussion* and *Information-transfer*). Drawbacks of the corpus include the fact that there are multiple speakers (it is therefore more difficult to capture contextual information) and sometimes its scheme does not map to the leaves of the ISO tree.

MapTask: This is a task-based dialogue corpus collected by the HCRC at the University of Edinburgh (Anderson et al., 1991). Dialogues involve two participants, one with an empty map and one with a route-marked map which must instruct the other speaker to draw the same route. The corpus was chosen due to its abundance of *Action-discussion* interactions (more than 30% of the overall corpus), which are often lacking in other corpora.

VerbMobil: This is a collection of task-based dialogues released in 1997 (Burger et al., 2000). A subset of these dialogues is annotated with DAs (Alexandersson et al., 1998). The scenario involves two speakers, which play respectively the roles of a travel agent and of a client. The client usually provides a set of constraints and requests to be satisfied, while the traveling agent has to ask questions and provide information in order to satisfy the client's requests. Interactions happening within the VerbMobil 2 corpus closely resemble those usually seen with personal assistants, with a user looking for the fulfillment of a task and a serving agent interacting with the user to solve his/her issues making it an appealing addition to our training set.

BT Oasis: The BT Oasis corpus is a collection of task-based conversations involving personal assistance for clients of the British Telecom services (Leech and Weisser, 2003). The conversations are human-to-human, and usually involve a user who has a problem to solve and an assistant who helps the user solving his issues. The BT Oasis corpus was chosen as part of the training set for its interesting scheme, called SPAAC (Speech Act Annotation scheme for Corpora), which is easily mappable to the ISO standard due to its clear separation of grammatical and illocutionary act.

3.2 Test sets

DialogBank (DB): the DialogBank (Bunt et al., 2016) is a corpus[2] annotated with ISO 24617-2 which currently contains 15 English dialogues: 3 from MapTask and 3 from TRAINS (Traum, 1996) (both task-based), 5 from DBOX (games collected in a Wizard-of-Oz fashion) and 4 from Switchboard (open-domain human-human conversation). Overall there are 1,596 DAs. The corpus currently represents the only publicly available resource manually annotated using the ISO standard.

Common Alexa Prize Conversations (CAPC): The CAPC corpus (Ram et al., 2017) is a dataset of 3,764 anonymised individual user turns pooled from different users interacting with all socialbots participating in the Alexa Prize. We have extracted a balanced subset of 458 turns and have annotated it with DAs from our adapted version of the ISO standard by 3 annotators, with an inter-annotator agreement of $\kappa = 0.82$. CAPC exemplifies frequent user interaction data not biased by the interaction with one socialbot in particular. Another advantage of this dataset is that it is balanced across different DA categories. One drawback is that no interaction context (previous DA) is available for the individual turns.

Socialbot Logs (S-Logs): S-Logs is a dataset of 13 open-domain conversations that different native American English speakers had with one of the socialbots of the Alexa Prize Challenge 2017. Overall this dataset contains 310 machine DAs and 165 user DAs. Two annotators tagged this dataset with DAs from our adapted version of ISO 24617-2, with an inter-annotator agreement of $\kappa = 0.81$. While we have annotated both machine and user turns, we test only on the latter and exploit machine turns as features for our classification experiments.

4 Methodology

4.1 Preprocessing

Before mapping the DA schemes of the corpora to the ISO subset scheme, a series of preprocessing steps have been performed to obtain a uniform training resource with the same surface text features as the testing corpora, since in S-Logs data, user input is lowercased and the punctuation is limited to apostrophes. More specifically, the text has been lowercased (including any named entity appearing in the original transcription and excluding the 'I' pronoun), punctuation has been removed (except for the apostrophe character in contracted expressions like "let's" and "can't") and any special characters have been deleted from the utterances. Moreover, any information regarding prosody has been removed, since this feature is not available in our test sets.

For experiments on the SWDA DAMSL corpus we recreated the same setting described in (Stolcke et al., 2000), using the same train and test set and preprocessing the corpus in the same way following the WS97 manual annotator guidelines (Jurafsky, 1997).

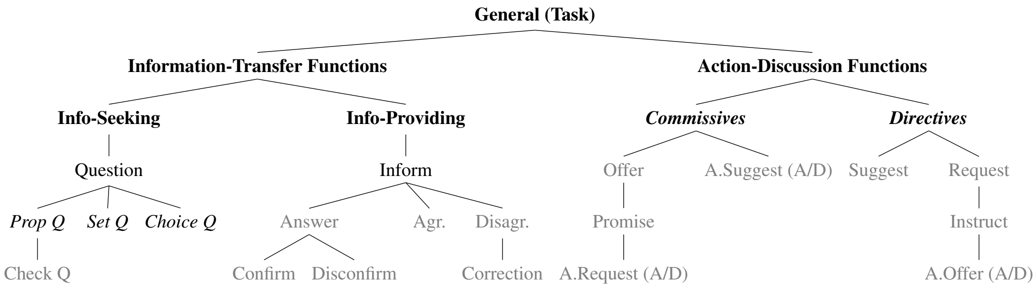

Figure 2: Communicative functions from ISO 24617-2 General purpose (Task) dimension. The nodes of the taxonomy that are not considered are grayed out.

S-scheme	**ISO 24617-2**
Social Obligation Management	
Salutation	Greeting, Goodbye, Self-Intro
Apology	Apology, Accept Apology
Thanking	Thanking, Accept Thanking
Feedback	
Feedback	Auto-Feedback (all), Allo-Feedback (all)

Table 1: Our scheme (S-scheme) compared to the corresponding ISO 24617-2 scheme for the SOM and Feedback dimensions

4.2 Dialogue Act scheme and mapping

The Socialbot scheme (S-scheme), the DA scheme used during the classification experiments, is a subset of the official ISO standard. Only three dimensions out of the official eight defined in the standard are considered (*Task*, *Social Obligation Management* and *Feedback*), and some of the communicative functions are generalized with an higher level of the tree.

Figure 4.2 shows the labeled subset of the ISO standard taxonomy for the General-purpose functions (i.e. functions independent from any given dimensions), while table 1 shows the correspondence for dimension-specific functions. The main difference between the DA scheme labeled in this work and the complete ISO taxonomy is the lack of further specification for the *Inform*, *Commissive* and *Directive* tags. This is due to the fact that most of the DA schemes used when building the training set do not provide contextual information detailed enough to label these tags accurately. Moreover, there is confusion and discrepancies about when these contextual DA should be used, even in the official ISO guidelines. Indeed, in (Fang et al., 2012), which provides the official mapping from Switchboard to the ISO standard, it is reported that some contextual DA tags (for example *other_answer*) do not have a direct mapping to the standard. This becomes even more problematic considering that among the training resources there are corpora like AMI, MapTask or VerbMobil, which label answers as Informs, which would make training data for this class extremely noisy. A similar argument can be raised on the lower leaves of the *Directive* and *Commissive* nodes, some of which are not labeled even in the very detailed SWDA taxonomy. Mapping of the available corpora to this scheme was done according to the available documentation in literature.

For the Switchboard corpus, a detailed mapping is provided in (Fang et al., 2012), which was followed exactly for the supported dimensions/communicative functions. For MapTask and AMI, there is already research highlighting similarities and differences between their schemes and the ISO standard one (Petukhova et al., 2014b). These results do not provide an exact mapping between the two schemes, which in some cases is impossible: for example the AMI *Elicit-inform* tag is the equivalent of ISO 24617-

[2] `https://dialogbank.uvt.nl/`

DA	**SWDA**	**MapTask**	**VerbMobil**	**Oasis BT**	**AMI**	**DB**	**CAPC**	**S-Logs**
Semantic Dimensions								
General (Task)	83,652	15,054	5,330	2587	1,523	1035	442	142
Social OM	2,866	0	384	588	10,039	21	16	7
Feedback	39,866	5,070	2,768	1,172	31,985	407	0	16
Total*	126,384	20,508	8,482	2,381	43,547	855	329	109
% of Corpus	79%	100%	72%	58%	74%	100%	100%	100%
General (Task) Dimension								
Commissives	63	-	7	25	1,523	57	20	1
Directives	7	4,075	2,911	181	10,039	131	93	32
Inform	75,667	4,860	-	1,648	33,403	652	105	91
Prop. Question	1,986	583	-	492	-	61	68	12
Set Question	5,506	1,692	-	241	-	134	149	6
Choice Question	423	-	-	-	-	8	7	-
Total*	83,652	11,210	2,918	2,587	44,965	1,035	442	142
% of Corpus	57%	30%	32%	35%	34%	N/A	N/A	N/A
Social Obligations Management								
Salutation	2,711	-	340	231	-	13	6	2
Apology	75	-	-	44	-	6	3	4
Thanking	80	-	44	193	-	2	7	1
Total*	2,866	0	384	468	2,201	21	16	7
% of Corpus	2%	0%	2%	8%	0%	N/A	N/A	N/A
Feedback								
Total	39,886	5,070	2,768	1,172	31,985	407	-	16
% of Corpus	79%	100%	72%	58%	74%	N/A	N/A	N/A

Table 2: Dialogue Act category counts across the considered corpora for different levels of the taxonomy. *Percentages of corpora* indicate the percentage of data available for the particular level in the corpus. $*$ It is frequently the case that DA tags do not map to any leaf-node, e.g. VerbMobil for Task dimension and AMI for Social Obligations Management.

2's *Question*, but will not map to any specific question type (*SetQ*, *PropQ*, *ChoiceQ*, etc.). Utterances whose tags cannot be directly mapped to the ISO scheme were dropped and do not appear in the training set.

Since there is no available literature on mapping the VerbMobil 2 and BT Oasis corpora to the ISO standard, a specific mapping was designed from scratch by drawing inspiration from the approaches available on other corpora.

Table 2 presents counts of the DAs after mapping to our scheme, across all training and testing corpora. As mentioned in the paper, the corpora present quite imbalanced distributions of DA categories.

5 Experiments and Results

Since, to the best of our knowledge, there is no established state of the art on DialogBank – the only corpus manually annotated following the ISO 24617-2 scheme – we first establish the tagging methodology on the SWDA corpus using the DAMSL 42 tag set and compare it to the state of the art (Ji et al., 2016). Then, the feature set and the parameters of the best performing models are used for the training of the DA tagger on the aggregate dataset, considering some of the semantic dimensions and the communicative functions of the ISO 24617-2 . The models are then evaluated on the DialogBank and open-domain human-machine data from Amazon Alexa Prize Challenge. McNemar's test (McNemar, 1947) for statistical significance has been used to analyze whether introduced features give a significant contribution to the overall performance.

5.1 Experiments on SWDA

Prior to training the classification models, the SWDA (Jurafsky, 1997) utterances are preprocessed following (Stolcke et al., 2000). The dataset is split into training (1,115 dialogues) and test set (19 dialogues) following the same paper, and the remaining 21 dialogues are used as development set to tune the C parameter of Support Vector Machines (SVM) (Vapnik, 1995). For the experiments, we used the SVM

Features	Acc.	Features	Acc.
BL: Majority	31.5		
HMM (Stolcke et al., 2000)	71.0	1-2-grams + PREV	74.6 *
SVM (Quarteroni and Riccardi, 2010)	72.4	1-2-grams + PREV + POS	74.6
DrLM (LSTM) (Ji et al., 2016)	77.0	1-2-grams + PREV + I-POS	76.2 *
1-grams	71.2	1-2-grams + PREV + I-POS + DEP	76.0
1-2-grams	71.7	1-2-grams + PREV + I-POS + I-DEP	76.1
1-2-3-grams	71.4	1-2-grams + PREV + I-POS + WE	**76.7** *

Table 3: Classification accuracy of the different feature combinations on the SWDA test set. The best results are highlighted in bold. The results that are significantly better are marked with *.

implementation of scikit-learn (Pedregosa et al., 2011) with linear kernel (i.e. its *liblinear* (Fan et al., 2008) wrapper).

The results of the experiments on SWDA are presented in Table 3. The performances are on the SWDA test set with the SVM C parameter set to 0.1, with respect to the best results on the development set. It is worth mentioning that tuning the C parameter boosts the performance on the development set by 2 points.[3] For comparison, the table also includes majority baseline, the results from (Stolcke et al., 2000), the SVM results from (Quarteroni and Riccardi, 2010), and the state-of-the-art results from (Ji et al., 2016) that were achieved using deep learning methods.

Following the previous studies on SWDA (Stolcke et al., 2000; Quarteroni and Riccardi, 2010), we experiment with n-grams (unigrams, bigrams, and trigrams) and previous DA tag features. We do not consider the unit length feature from (Quarteroni and Riccardi, 2010), since classification instances in the SWDA scheme and ISO 24617-2 are different (slash unit vs. functional unit). The results are reported in Table 3; since the results reported were obtained with SVM C parameter set to 0.1, they are higher than the ones reported in (Quarteroni and Riccardi, 2010): e.g. for 1-2-grams 70.0 vs. 71.7.

The first observation is that the addition of the previous DA significantly improves the performance. Addition of part-of-speech tags does not yield any improvement; however, when POS-tags are indexed with their positions in an utterance, accuracy is significantly improved and rises to 76.2. Addition of dependency relations (both with and without indexing with their position) does not improve the performance. Addition of the averaged pre-trained word-embedding vectors (from Google News) to the model with indexed POS-tags, however, rises the accuracy to 76.7. The model with word embeddings comes 0.3 short of the state-of-the-art results reported in (Ji et al., 2016).

5.2 Experiments on Aggregate ISO-standard Data

The methodology established on SWDA is applied to training the ISO 24617-2 subset models using the aggregate data set. Since in ISO 24617-2 annotation scheme DAs consist of semantic dimensions and communicative functions, the utterances are first classified into the considered semantic dimensions – general, social obligations management (SOM), and feedback. Then, we experiment with the Task dimension, reporting the results without error propagation from the previous step, in order to give the reader a clearer understanding of the current classification capabilities when restricting interactions with the system to general communicative functions.

5.2.1 Semantic Dimension Classification

The results of the binary dimension classification models on the test sets – DialogBank (DB), CAPC, and S-Logs – are reported in Table 4. The CAPC corpus consists of isolated utterances; consequently, the *Feedback* dimension is not present. On DB and S-Logs, on the other hand, the *Feedback* dimension yields the lowest accuracy in comparison to General and *SOM* dimension communicative functions. Low performances on the *Feedback* dimension could be explained by the fact that the training data mostly contains *Allo-feedback* and lacks *Auto-feedback* and *Feedback elicitations*, which are present in DB.

[3] From 73 ($C = 1.0$) to 75 ($C = 0.1$) for the model trained on unigrams, bigrams and previous DA-tag.

Dimension	DB	CAPC	S-Logs
General	73.3	83.0	80.2
SOM	78.1	90.7	86.6
Feedback	56.3	–	71.3
Overall	68.4	83.3	79.4

Table 4: Classification accuracies of the binary semantic dimension models: General, SOM and Feedback. The CAPC corpus does not contain Feedbacks, therefore results for this dimension are not reported.

Features	DB	CAPC	S-Logs	Features	DB	CAPC	S-Logs
BL: Majority	53.4	22.9	63.4				
1-2-grams	64.2	71.2	78.7	+ I-DEP	67.1*	74.3	82.3
+ PREV	64.3	70.7	81.6*	+ WE	65.2	74.8*	82.0
+ I-POS	65.8*	73.8*	82.2	+ I-DEP + WE	66.6	75.1	81.8

Table 5: Accuracies of the feature combinations on the general-purpose communicative functions on the test sets. The best results are marked in bold, and statistically significant differences with *.

5.2.2 Communicative Function Classification

The utterances are further classified into communicative functions of the General (Task) dimension, using the methodology established on SWDA, i.e. the same hyper-parameter settings ($C = 0.1$) and features. However, since models with dependency relations do not yield statistically significant differences, they are also considered. The results of the models on the test sets are reported in Table 5. The behavior of the models trained with various feature combinations is in-line with the SWDA experiments: the addition of the previous DA tags and part-of-speech tags indexed with their positions in a sentence improves the performance. Different from the SWDA, the addition of the indexed dependency relations improves the performance on the test sets. In the case of DialogBank and CAPC, their contribution is statistically significant. Additionally, unlike for SWDA, the addition of word embeddings with and without index dependency relation (I-DEP) does not produce significant improvements for all but CAPC. Consequently, the model trained on 1-2-grams, previous DA tags, indexed POS-tags and dependency relations is chosen for the ablation study.

5.2.3 Corpora Combinations

The aggregation of all the corpora mapped to our subset of ISO 24617-2 is not necessarily the best one, as the distributions of DA categories varies from corpus to corpus. Consequently, we also present results on the test sets for the models trained solely on SWDA and AMI; as well as perform an ablation experiment removing one corpus at a time. The best performing model from the previous subsection (1-2-grams, previous DA-tag, indexed POS-tags, and indexed dependency relations) is used for the study. The results of these experiments are reported in Table 6.

While the best results for Dialog Bank are achieved considering all the corpora, for CAPC the best results are achieved by removing MapTask. For S-Logs, on the other hand, the best performing corpora combination is all except VerbMobil. However, the performance differences from the models trained on all corpora are not statistically significant. Training DA taggers solely on SWDA and AMI – the largest and the most diverse corpora – yields performances inferior to the combination of all the corpora. From the table, we can also observe that these two corpora – SWDA and AMI – contribute most to the performance, as removing them affects the performance the most. On the other hand, removing the smaller datasets – BT Oasis, MapTask, and VerbMobil – affects the performance less.

6 Conclusions

We have presented an effective methodology for corpora aggregation for domain-independent Dialogue Act Tagging on a subset of the ISO 24617-2 annotation. We have also reported an accurate evaluation

Dataset	DB	CAPC	S-Logs	Dataset	DB	CAPC	S-Logs
ALL	**67.1**	74.3	82.3	- AMI	59.7	73.7	71.3
				- SWDA	60.2	68.3	77.5
SWDA only	57.9	71.3	53.5	- Oasis BT	66.1	74.2	81.8
AMI only	53.2	39.8	61.6	- MapTask	66.8	**74.6**	80.5
				- VerbMobil	66.5	74.0	**82.6**

Table 6: Accuracies of the corpora combinations on the test sets – Dialog Bank (DB), CAPC, and S-Logs.

of our approach on both in-domain and out-of-domain datasets, proving that the described DA tagging technique is indeed independent from the underlying scheme and task of the annotated corpora. Finally, the machine learning technique used for DA tagging was tested on a popular DA tagging task (the Switchboard corpus), obtaining very close to state-of-the-art results.

This work represents one of the first attempts to use an ISO compliant DA scheme for a real-life application, as well as one of the first structured approaches for evaluation of dialogue resources annotated with this taxonomy.

Research on available training resources is one of the first things to look forward to, since the current data proved to be effective, but also presented numerous drawbacks (lack of adequate coverage for the Feedback dimension, imbalanced DAs, lack of context-aware communicative functions). We plan to make our resource continue to grow in the future by adding and mapping additional corpora, such as the MRDA (Shriberg et al., 2004) corpus.

7 Acknowledgments

This project was partially funded by Amazon Alexa Prize 2017 research grant.

References

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. *Dialogue acts in Verbmobil 2*. DFKI Saarbrücken.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

John Langshaw Austin and JO Urmson. 1962. *How to Do Things with Words. The William James Lectures Delivered at Harvard University in 1955.[Edited by James O. Urmson.].* Clarendon Press.

Kevin K Bowden, Jiaqi Wu, Shereen Oraby, Amita Misra, and Marilyn Walker. 2017. Slugbot: An application of a novel and scalable open domain socialbot framework. *Alexa Prize Proceedings*.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an iso standard for dialogue act annotation. *Seventh conference on International Language Resources and Evaluation (LREC'10).*

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Chengyu Fang. 2016. The dialogbank. In *LREC*.

Harry Bunt. 1999. Dynamic interpretation and dialogue theory. *The structure of multimodal dialogue*, 2:1–8.

Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Susanne Burger, Karl Weilhammer, Florian Schiel, and Hans G Tillmann. 2000. Verbmobil data collection and annotation. In *Verbmobil: Foundations of speech-to-speech translation*, pages 537–549. Springer.

Jean Carletta. 2006. Announcing the ami meeting corpus. *The ELRA Newsletter 11(1), January-March, p. 3-5*.

Alessandra Cervone, Giuliano Tortoreto, Stefano Mezza, Enrico Gambi, and Giuseppe Riccardi. 2017. Roving mind: a balancing act between open–domain and engaging dialogue systems. In *Alexa Prize Proceedings*.

Shammur Absar Chowdhury, Evgeny A Stepanov, and Giuseppe Riccardi. 2016. Transfer of corpus-specific dialogue act annotation to iso standard: Is it worth it? In *LREC*.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Alex C. Fang, Jing Cao, Harry Bunt, and Xiaoyue Liu. 2012. The annotation of the Switchboard Corpus with the new ISO standard for dialogue act analysis. In *Workshop on Interoperable Semantic Annotation*.

Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mary Ostendorf, Yejin Choi, and Noah A. Smith. 2017. Sounding board – university of washington's alexa prize submission. In *Alexa Prize Proceedings*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Narendra Gupta, Gokhan Tur, Dilek Hakkani-Tur, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert. 2006. The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *SIGDIAL Conference*, pages 263–272.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of NAACL-HLT*, pages 332–342.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function. *Annotation, Technical Report, 97-02, University of Colorado, CO, USA*.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with crf. *arXiv preprint arXiv:1709.04250*.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In *Procs. of the 2003 Corpus Linguistics Conference, pp. 441Y446. Centre for Computer Corpus Research on Language Technical Papers, Lancaster University*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, Jun.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Anna Schmidt, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, et al. 2014a. The dbox corpus collection of spoken human-human and human-machine dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC" 14)*, number EPFL-CONF-201766. European Language Resources Association (ELRA).

Volha Petukhova, Andrei Malchanau, and Harry Bunt. 2014b. Interoperability of dialogue corpora through ISO 24617-2-based querying. In *LREC*.

Patti J Price. 1990. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Silvia Quarteroni and Giuseppe Riccardi. 2010. Classifying dialog acts in human-human and human-machine spoken conversations. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pages 2514–2517.

Silvia Quarteroni, Alexei V Ivanov, and Giuseppe Riccardi. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5596–5599. IEEE.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2017. Conversational ai: The science behind the alexa prize. In *Alexa Prize Proceedings*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3).

David Traum. 1996. Conversational agency: The trains-93 dialogue manager. In *In Susann LuperFoy, Anton Nijhholt, and Gert Veldhuijzen van Zanten, editors, Proceedings of Twente Workshop on Language Technology, TWLT-II*. Citeseer.

V.N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 78–83. IEEE.

Xuesong Yang, Yun-Nung Chen, Dilek Hakkani-Tür, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng. 2017. End-to-end joint learning of natural language understanding and dialogue manager. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5690–5694. IEEE.

A Appendix A. Dialogue Acts mappings

ISO	SWDA	MapTask	VerbMobil	Oasis BT	AMI
Task					
Inform	Statement-non-opinion, Statement-opinion, Rhetorical-question, Statement expanding y/n answer, Hedge	explain, clarify	–	Inform	Inform
ChoiceQ	Or-question Or-clause	–	–	–	–
SetQ	Wh-question Declarative wh-question	query_w	–	q_wh	–
PropQ	Yes-no-question, Backchannel in question form, Tag-question, Declarative Yes-no-question	query_yn	–	q_yn	–
Commissive	Offer, Commit	-	Offer, Commit	Offer	Offer
Directive	Open-Option	Instruct	Request (all), Suggest	Suggest, imp	Suggest, Elicit-offer
Social Obligation Management					
Thanking	Thanking, You're-welcome	–	–	thank	–
Apology	Apology Downplayer	–	–	pardon regret	–
Salutation	Conventional-closing	–	–	bye, greet	–
Feedback					
Feedback	Signal-not-understanding, Acknowledge (backchannel), Acknowledge answer, Appreciation, Sympathy, Summarize/ reformulate, Repeat-phrase	Acknowledge	Feedback (all)	ackn	Backchannel

Table 7: Mapping for Dialogue Acts from each individual corpus to ISO DA-tags.