# A Multi-Dimensional, Cross-Domain and Hierarchy-Aware Neural Architecture for ISO-Standard Dialogue Act Tagging

**Stefano Mezza, Wayne Wobcke and Alan Blair**
School of Computer Science and Engineering
University of New South Wales, Sydney NSW 2052, Australia
{s.mezza,w.wobcke,a.blair}@unsw.edu.au

## Abstract

Dialogue Act tagging with the ISO 24617-2 standard is a difficult task that involves multi-label text classification across a diverse set of labels covering semantic, syntactic and pragmatic aspects of dialogue. The lack of an adequately sized training set annotated with this taxonomy is a major problem when using the standard in practice. In this work, we propose a neural architecture to increase classification accuracy, especially on low-frequency fine-grained tags, on a subset of the ISO 24617-2 taxonomy. Our model takes advantage of the hierarchical structure of the ISO taxonomy and utilises syntactic information in the form of Part-Of-Speech and dependency tags, in addition to contextual information from previous turns. We train our architecture on an aggregated corpus of conversations from different domains, which provides a variety of dialogue interactions and linguistic registers. Our approach achieves state-of-the-art tagging results on the DialogBank benchmark data set, providing empirical evidence that this architecture can successfully generalise to different domains.

## 1 Introduction

Language understanding is a fundamental component of any conversational system, as it impacts its abilities to correctly recognise a user's communicative functions and act accordingly. Dialogue Act (DA) tagging is a crucial step of this understanding process, particularly in an open-ended conversational setting, as it informs the system on the users' beliefs, desires, intentions and actions.

Table 1 shows an excerpt of a conversation from the Mastodon corpus annotated with Dialogue Act tags. Note that some tags (e.g. *Task:Answer* or *Task:Agreement*) are contextual and depend on the tagging of previous utterances. The annotation also reflects the multi-dimensional nature of the DA tags, categorised as *Task*, *Social* or *Feedback*. In general, an utterance may have multiple tags, even of the same dimension.

| Utterance | DA Tags |
|---|---|
| A: ask anything you'd like | Task:Directive |
| B: thanks for the interest | Social:Thanking |
| B: when a girl keeps blinding you with the reflection of the sun is she signalling that she wants to hold hands ? | Task:InfoQuestion |
| A: only if the flash pattern is .. - . or maybe ... - | Task:Answer |
| A: Deleted my Facebook account a few days ago and I never felt so free in my entire life. | Task:Inform |
| A: Now I just have to encourage my closest friends to do the same | Task:Commissive |
| B: It shouldn't be that hard. They are as tired of social media as I am . | Task:Inform |
| A: Yes ! I don't get it . Everyone I talk to about Facebook– EVERYONE - - hates it , but none of them will take action . | Task:Agreement |

Table 1: An example dialogue from the Mastodon corpus annotated with the ISO 24617-2 taxonomy.

Early dialogue applications usually adopted a list of mutually exclusive and task-specific DA tags which represented the different functions that the system performed, acting essentially as *intent labels*. These taxonomies were also *one-dimensional*, featuring mutually-exclusive tags which did not account for the complexity of the dialogues. The following example from Bunt (2006) clarifies the importance of multi-dimensionality in DA tagging:

S: Can you tell me what time is the first train to the airport on Sunday morning?

A: On Sunday morning the first train to the airport is at 5.32.

S: Thank you!

According to Bunt (2006), the third utterance has two separate communicative functions, as the speaker S is expressing gratitude towards the addressee A (*Social* dimension), while at the same time informing them on their understanding of the train schedule (*Feedback* dimension).

In addition to this, DA taxonomies have also typically lacked a hierarchical organisation of tags, which makes it difficult for a classifier to capture the high-level mutual relationships between dialogue tags (Soria and Pirrelli, 2003).

In an attempt to address these problems, an official ISO standard taxonomy, ISO 246170-2, was introduced in Bunt et al. (2012): this taxonomy is domain-independent, hierarchical and multi-dimensional, and well-suited for open-ended Natural Language Understanding. However, much work in dialogue systems still uses other taxonomies, in part we believe because of the lack of an adequately-sized data set annotated with the standard, which makes it difficult to train a classifier for the taxonomy. Some authors have proposed automated mappings of old resources to the new ISO standard; however, these works are still limited in scope, focusing either on heavily imbalanced data sets (Bunt et al., 2017) or subsets of the ISO 24617-2 taxonomy that are insufficient for real-life conversational scenarios (Mezza et al., 2018). DA tagging with the ISO taxonomy is also an intrinsically difficult task, as it requires handling multiple different dimensions and a collection of different fine-grained tags which differ in semantic, syntactic and contextual aspects. The open-ended nature of the taxonomy provides an additional challenge, as most existing DA-annotated resources tend to have a bias towards specific topics or discussion styles, which hinders the model's capabilities to generalise to unseen conversations.

In this work, we introduce a neural architecture optimised for DA tagging with a subset of the ISO standard taxonomy. Our model combines syntactic, semantic and contextual information and leverages the hierarchical dependencies across labels to improve the classification accuracy, especially on low-frequency fine-grained tags. We combine existing DA-annotated data sets and map them to a subset of the ISO 24617-2 taxonomy to obtain an adequately-sized training set, taking advantage of existing mappings in the literature (Mezza et al., 2018) and novel conversational resources such as Mastodon (Cerisara et al., 2018) and DailyDialog (Li et al., 2017), whose taxonomies easily map to the ISO standard. We also experiment with the addition of online discussions and debates data from the Internet Argument Corpus v2.0 (Abbott et al., 2016; Walker et al., 2012), in order to increase the system's understanding of opinionated and contextual tags. Experimental results show that our approach achieves state-of-the-art classification accuracy on the DialogBank (Bunt et al., 2016) test set. We also provide additional experiments that delve into the details of the training process, including ablation studies and an analysis of the extent to which the different corpora that we utilised contribute to the network's performance. Finally, we share our code and our mapped data set [1], in order to share these resources with the research community and hopefully encourage further research on Dialogue Act classification.

## 2 Related Work

The concept of Dialogue Act (DA) has its roots in the seminal works by Austin (1975) and Searle (1965), who established the theoretical foundations of the Speech Act theory. A *speech act* captures an utterance at the level of its illocutionary force. Many subsequent works started referring to speech acts in a conversational setting as *Dialogue Acts*, and investigated possible taxonomies of Dialogue Acts. These early taxonomies were flat (there was no distinction between coarse-grained and fine-grained tags), mono-dimensional (each and every utterance had only one Dialogue Act tag) and usually task-specific rather than domain-independent. Examples of these early taxonomies include the DAMSL taxonomy (Allen and Core, 1997), which was used for the annotation of the Switchboard and MRDA conversational corpora, the HCRC coding manual (Anderson et al., 1991), which was used to annotate the Maptask corpus, and the VerbMobil annotation scheme (Jekat et al., 1995), which was used for the annotation of the homonymous corpus.

An interest in formalising these taxonomies into a more rigid theoretical framework arose in the early 2000s, with Traum (2000) analysing existing DA taxonomies and investigating a rigorous definition of Dialogue Acts. Bunt (2005) provided one of the first formal definitions of Dialogue Act as "a unit in the semantic description of communicative behaviour, produced by a sender and directed at an addressee, specifying how the behaviour is intended to influence the context through understanding of the behaviour". The authors combined existing taxonomies such as DAMSL and DIT (Bunt, 1989) into a new taxonomy called DIT++ (Bunt, 2009), which aimed at being a truly open-ended, domain-independent and theoretically sound tax-

[1] https://github.com/coling22tagger/DialogueActTagger

onomy. The fifth version of the DIT++ taxonomy became the official ISO standard for Dialogue Act classification (Bunt et al., 2012). A potential advantage of using the standard is that multiple corpora can be used to construct a larger training set, or cross-domain training sets suitable to cover a wide rage of dialogue tasks; nonetheless, the ISO 24617-2 taxonomy has yet to be fully adopted, with many works still using DAMSL as their target taxonomy (Raheja and Tetreault, 2019; Cervone et al., 2018), or introducing entirely novel taxonomies custom-tailored for specific tasks (Paul et al., 2019; Yu and Yu, 2019). While this is partly due to the complexity of the standard, we believe a significant obstacle is the lack of adequately-sized data sets to train a classifier. Some efforts have been made to convert existing resources to the new taxonomy: Fang et al. (2012) proposed an automated, albeit partial, mapping of the DAMSL taxonomy to the ISO standard, Mezza et al. (2018) extended their work providing partial mappings for the AMI, MapTask, Oasis and VerbMobil taxonomies, and Ribeiro et al. (2020) provides a mapping from the LEGO annotation scheme to the ISO one. There are some planned corpora entirely annotated with the ISO standard, such as ADELE (Gilmartin et al., 2018) or DBOX (Petukhova et al., 2014), which may become valuable tools to work with the taxonomy. However, these resources are still not publicly available at the time of writing. Some interesting corpora were released in recent years which, while not being entirely ISO compliant, adopted DA taxonomies which can be easily mapped to the standard. These include DailyDialog (Li et al., 2017), Mastodon (Cerisara et al., 2018) and MIDAS (Yu and Yu, 2019) among others.

Automatic DA tagging was initially formalised as a text classification task by Stolcke et al. (2000), who presented a Hidden Markov Model for the classification of the Switchboard data set. Since then, many different approaches have been proposed to tackle the task, including rule-based systems (Lendvai et al., 2003), Conditional Random Fields (Quarteroni et al., 2011; Zhou et al., 2014) and Support Vector Machines (Mezza et al., 2018). More recent works shifted their focus to Artificial Neural Networks, which have been proved to be very effective for text classification tasks; many of these models are built around Bi-LSTM/GRU cells with CRF as a top layer, due to this architecture's ability to capture long-term contextual dependencies in dialogue (Kumar et al., 2018; Chen et al., 2018). Many architectures also rely on transformers, usually combined with pre-trained sentence embeddings (Yu and Yu, 2019) or some form of attention mechanism (Raheja and Tetreault, 2019). The vast majority of these models are designed for flat hierarchies of tags, with little to no emphasis on the multi-dimensional nature of Dialogue Acts. Anikina and Kruijff-Korbayova (2019) annotated a subset of the TRADR corpus of robot-assisted disaster response team communications with three dimensions of the ISO standard (*General (Task)*, *Social* and *Turn Management*) and trained various neural classifiers for the task, including CNN, LSTM and FFN; while their work does take advantage of the multi-dimensional aspect of the ISO taxonomy, the scope of their research is limited by the size of the resource and the emphasis on the disaster response domain.

A number of recent works started taking advantage of the hierarchical structure of ISO communicative functions: Wang et al. (2021) introduced a hierarchical neural model for one-dimensional DA tagging (they only consider the Social and Task dimensions of the standard and combine them into a single core dimension), while Ribeiro et al. (2019) proposes a hierarchical and multi-dimensional approach for the Spanish corpus DIHANA. Blache et al. (2020) apply a number of statistical machine learning algorithms, such as *XGBoost* and *Random Forests*, to annotate French medical data with a subset of the ISO standard; their approach separates the classification into two hierarchical steps to increase the accuracy of the model. Mezza et al. (2018) proposed a multi-dimensional and domain-independent approach to DA tagging, and also investigated hierarchical DA tagging through a tree-like structure of SVM classifiers; however, the model was limited in scope and accuracy and did not take into account contextual tags such as *Answer*, *Agree/Disagree*, etc. Ribeiro et al. (2022) utilised an end-to-end hierarchical network with cascading outputs and maximum a posteriori path estimation to classify all the layers of the *General (Task)* semantic dimension; while their architecture handles the whole taxonomy of *Task* communicative functions, it lacks support for additional dimensions of the standard such as *Social*, *Turn Management*, etc. Their model also fails to capture the domain-independent nature of the taxonomy, as its performance degrades with the addition of

out-of-domain data such as the conversations from the LEGO-ISO corpus.

Multi-label Hierarchical Text Classification has been successfully addressed for other text classification tasks in the literature, such as fine-grained Sentiment Analysis (Tai et al., 2015) or Topic Classification (Zhou et al., 2020). The latter work proposed two different models to solve hierarchical, multi-dimensional topic classification on news articles; we adapt a similar approach to include prior information on the hierarchical correlation among labels in our model.

## 3 Methodology

### 3.1 Task Definition

A *dialogue* $D$ is defined as a sequence of *dialogue turns* $T_1, \ldots, T_n$, with each turn performed by an individual speaker. Each turn consists of a sequence of *utterances* $u_1, \ldots, u_m$ performed by one of the speakers, with each utterance representing one of the *functional segments* of the turn (i.e. "a minimal stretch of functionally relevant communicative behaviour" (Bunt et al., 2010)). We have a taxonomy of tags $t_1, \ldots, t_n$ arranged in a set of tree-like structures. Each tree corresponds to a *core dimension* (an aspect of utterance function), and groups together tags that correspond to *communicative functions* within the same dimension. Fine-grained DA tags are the leaves of the trees, while coarse-grained tags are the intermediate nodes. *Dialogue Act (DA) tagging* is the task of assigning one or more fine-grained tags $t_1, \ldots, t_k$ to each utterance in the dialogue. Similarly to Mezza et al. (2018), we have decided to adopt a subset of the ISO 24617-2 standard taxonomy (Bunt et al., 2012) for our classifier, since some of the fine-grained tags of the standard do not appear in any of our corpora. We consider three core dimensions of the standard, namely *Task*, *Social* and *Feedback*, and a total of 16 fine-grained DA tags. Figure 1 shows our complete taxonomy.

### 3.2 Data

There is a widely recognised shortage of conversational data annotated with the ISO 24617-2 standard. Researchers have worked around this issue by designing their own taxonomies (Paul et al., 2019), using older more widely supported taxonomies (Raheja and Tetreault, 2019) or converting existing resources via rule-based mappings (Mezza et al., 2018; Ribeiro et al., 2020). We followed the latter approach and converted a number of resources to our subset of the ISO standard. The resulting aggregated corpus, **General Dialogue Corpus (GDC)**, is a combination of the following corpora:

- The **Switchboard Dialog Act Corpus (SWDA)** (Jurafsky and Shriberg, 1997), a collection of 5-minute telephone conversations on provided topics such as child care, recycling, and news media, annotated with the DAMSL taxonomy. Conversations in the corpus focus on information exchange, with an abundance of *Info Providing* and *Info Seeking* dialogue acts. They also feature a high number of *Feedback* tags due to the nature of telephone conversations, which often need explicit feedback to signal understanding.

- The **ICSI Meeting Recorder Dialog Act (MRDA)** corpus (Shriberg et al., 2004), a collection of transcribed research meetings annotated with a slightly edited version of the DAMSL taxonomy. Similarly to SWDA, this corpus contains a majority of information exchange tags; however, as the conversations involve multiple participants in an academic environment, there is also a significant amount of *conversational structuring* tags (including *Feedback*) and a more formal linguistic register.

- The **DailyDialog** corpus, a human-written and manually labelled set of DA annotated conversations about the daily life of the participants. The corpus focuses on social interactions among human speakers, with a good balance of *Information-Transfer* and *Action-Discussion* tags.

- The **Mastodon** corpus (Cerisara et al., 2018), a Twitter-like corpus of conversation threads on an open-source social platform called Mastodon. It features a combination of information exchange and persuasive dialogue and is annotated with sentiment information and coarse-grained DA tags.

- The **Internet Argument Corpus v2 (IAC)** (Abbott et al., 2016; Walker et al., 2012) is a collection of corpora for research on political debate on Internet forums. We focus on the *4Forums* subset of the resource, which contains argumentative dialogue and features agreement/disagreement stance annotations.

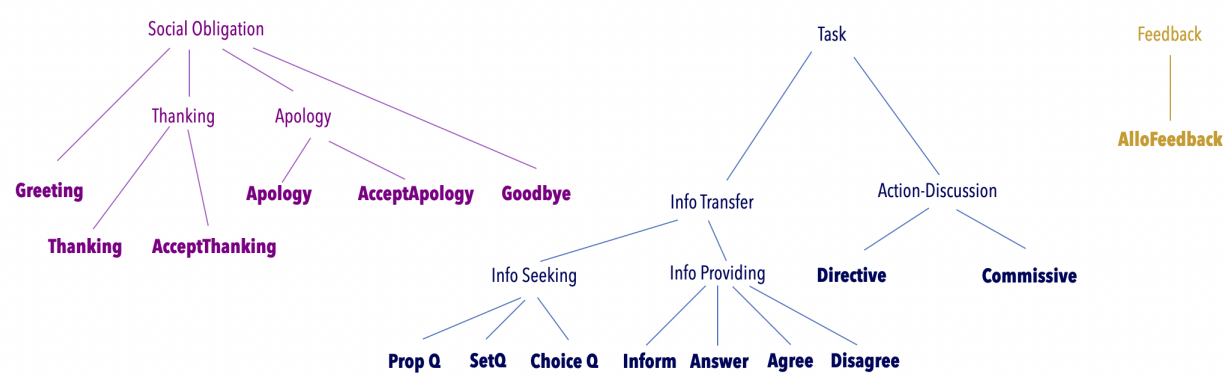


Figure 1: Our subset of the ISO 24617-2 standard for DA tagging.

We chose this particular collection of corpora to have a reasonably balanced distribution of tags across different core dimensions: SWDA and MRDA focus on information exchange and conversational structuring tags, DailyDialog has a high proportion of Action-Discussion tags (about 20% of the overall corpus), and Mastodon and IAC have an emphasis on opinionated, argumentative and persuasive dialogue. These corpora also offer a variety of linguistic registers: Switchboard and DailyDialog focus on everyday conversations with colloquial language, MRDA and IAC contain more formal conversations with a richer vocabulary and lexicon, and Mastodon features an abundance of Internet and chat slang.

We followed the mapping introduced in Fang et al. (2012) for the conversion of SWDA and MRDA taxonomies. The Mastodon corpus utilises a subset of the ISO standard, therefore we adopted the mapping suggested by the authors of the corpus (Cerisara et al., 2018). DailyDialog features coarse-grained DAs which directly map to coarse-grained tags in the General (Task) dimension of the ISO standard. Finally, we converted the *4Forums* subset of IAC by utilising the agreement and disagreement stance annotation to map responses to the *Agreement* and *Disagreement* tags of the standard. More specifically, we labelled responses with an agreement stance lower than -2.0 to *Disagreement*, responses with a stance higher than 2.0 to *Agreement* and all other responses to *Answer*. The train, test and validation splits of the GDC is a combination of all the splits of the included corpora. We utilised the default train, test and validation splits for the MRDA, DailyDialog and SWDA corpora, with the only variation being the removal of the SWDA conversations that appear in the DialogBank from the training and validation splits of the corpus. Given the large size of the Mastodon test split, we have elected to only use the first 500 utterances of the corpus for testing; we reserved 436 utterances as additional training data and 205 utterances for our validation split. Since the *4Forums* corpus of IAC does not have a default train-test split, we just divided the corpus manually and reserved 7847 responses for training, 638 responses for testing and the remaining 1497 for validation. We ensured that utterances belonging to the same conversation would be in the same split when dividing the corpora.

In addition to the test split of GDC, we also tested our model on the **DialogBank** corpus (Bunt et al., 2016), a collection of conversations from different corpora annotated with the ISO taxonomy by the authors of the standard; this is one of the few resources available that are manually annotated with the ISO taxonomy, and therefore constitutes a popular testing benchmark for the task.

### 3.3 Model

This section describes our model, **Dialogue Act Syntax and Hierarchy-aware Network (DASHNet)**. Figure 2 provides an overview of the main components of the network. DASHNet uses a triple input encoding mechanism: the *lexical encoder* encodes input tokens, the *syntax encoder* encodes syntactical information such as Part-of-speech (POS) tags and Dependency (DEP) tags, and the *context encoder* encodes contextual information from the previous speaker's last utterance. We have limited the context to a single previous utterance to test the hypothesis that some of the tags of the ISO 24617-2 taxonomy are contextual and directly

depend on the tagging of the previous utterance in the dialogue; moreover, some of our resources (such as the IAC) do not provide a context longer than one utterance. Given an input utterance as a sequence of tokens $Uj = t_1, \ldots, t_n$, the *lexical encoder* passes it through a pre-trained embedding layer and a bidirectional GRU layer to obtain the lexical encoding:

$$\vec{E_t} = Embedding(t_1, \ldots, t_n) = \vec{e_1^t}, \ldots, \vec{e_n^t} \quad (1)$$

$$\vec{H_L} = BiGRU(\vec{e_1^t}, \ldots, \vec{e_n^t}) \quad (2)$$

Similarly, grammatical features are encoded in the *syntax encoder* module with a linear layer replacing the pre-trained embeddings:

$$\vec{E_{pos}} = Linear(p_1, \ldots, p_n) = \vec{e_1^p}, \ldots, \vec{e_n^p} \quad (3)$$

$$\vec{E_{dep}} = Linear(d_1, \ldots, d_n) = \vec{e_1^d}, \ldots, \vec{e_n^d} \quad (4)$$

$$\vec{H_G} = \begin{pmatrix} BiGRU(\vec{e_1^p}, \ldots, \vec{e_n^p}) \\ BiGRU(\vec{e_1^d}, \ldots, \vec{e_n^d}) \end{pmatrix} \quad (5)$$

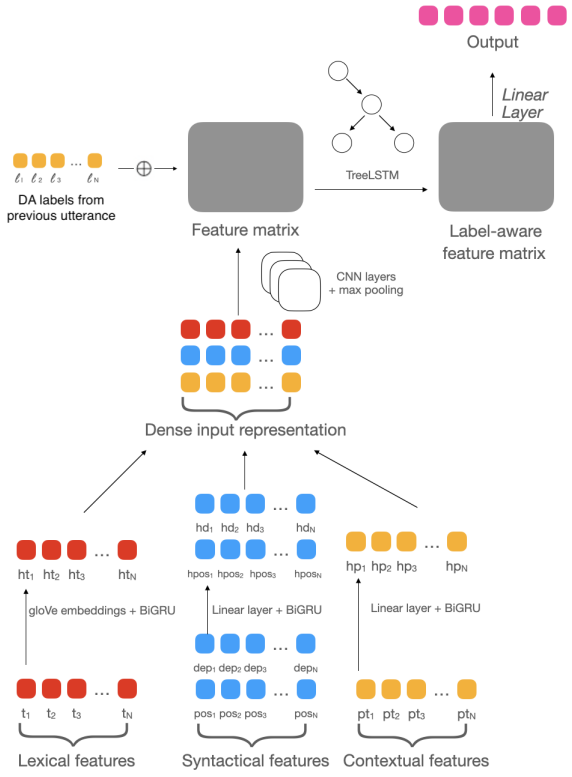


Figure 2: DASHNet architecture.

The *context encoder* input is just the *input encoding* from the previous sentence $U_{j-1}$. The input encoding for utterance $U_j$ will thus be

$$\vec{I_j} = \begin{pmatrix} H_{(L,j)} \\ H_{(G,j)} \\ I_{(j-1)} \end{pmatrix} \quad (6)$$

Multiple convolutional layers with different kernel sizes, followed by max-pooling, are used to extract relevant features from the input encoding $I_J$ obtaining a feature matrix $F_J$. An encoding of the list of DA labels from the previous speaker's last utterance is also concatenated to $F_J$.

Prior knowledge about the hierarchical correlation among labels is then embedded into the feature matrix. More specifically, we estimate the prior probabilities for each tag from the training data distribution as follows:

$$P(L_j|L_i) = \frac{N_j}{\sum_{k \in child(i)} N_k} \quad (7)$$

$$P(L_i|L_j) = 1.0 \quad (8)$$

where $P(L_j|L_i)$ denotes the probability of the fine-grained DA tag $j$ given the parent coarse-grained node $i$, $P(L_i|L_j)$ denotes the probability of the parent node $i$ given the child node $j$, $child(i)$ denotes the set of children nodes for tag $i$ and $N_k$ denotes the number of occurrences of tag $k$ in the training set. Since tags in the ISO 24617-2 taxonomy are arranged in a tree structure, the probability of a coarse-grained tag given the occurrence of any of its fine-grained children tags is always equal to 1. We compute these prior probabilities before training, and then encode them in the network through a Bidirectional Tree-LSTM. We use the implementation of BiTree-LSTM introduced in Zhou et al. (2020), which is itself based on the structure encoder presented in Li et al. (2018). Namely, the output from the CNN layers is then transformed through a linear layer to obtain a hidden label representation $l_i$ for each label in the taxonomy (including coarse-grained DA tags). The hidden state $h_k$ for DA tag $t_k$ is then computed as:

$$h_k = h_{k\downarrow} \oplus h_{k\uparrow} \quad (9)$$

$$\text{where} \quad h_{k\downarrow} = \sum_{i \in child(k)} P(L_i|L_k)\, h_i \quad (10)$$

$$\text{and} \quad h_{k\uparrow} = P(L_p|L_k) h_p \quad (11)$$

and $p$ represents the parent node for node $k$.

### 3.4 Experimental Setting

We trained our models on Google Colab Pro with CUDA GPU and High Memory settings for 100 epochs, with learning rate $\alpha = 1 \times 10^{-5}$, and use Adam optimiser with weight decay $w = 1 \times 10^{-4}$. We used pre-trained 300-dimensional GloVe embeddings and Kaiming uniform initialisation for weight initialisation (GloVe embeddings are used to facilitate comparison with previous work). The BiGRU layers for input representation have a hidden size of 128 nodes, and the node representation for DA labels in the Tree-LSTM structure is 300-dimensional. We use three CNN layers with kernel size 3, 4 and 5 respectively, with 100 filters each. Finally, we apply Dropout with probability $p = 0.3$ after each BiGRU layer, with probability $p = 0.5$ after the CNN layers and with probability $p = 0.1$ after the Tree-LSTM structure encoder. The values for the hyper-parameters of the network were chosen according to the average Macro-F1 score of the network on the validation split of the GDC corpus.

We extract Part-Of-Speech tags and dependency tags with the spaCy 3.1 Python library, which we also use to tokenise the input utterances. Since our mappings are partial, some of the utterances in the corpora could not be accurately annotated with any ISO communicative functions; these data points were annotated with coarse-grained DA tags and used as contextual features where appropriate (for example, tags *Inform* and *Question* from DailyDialog can be mapped to *Task:InfoProviding* and *Task:InfoSeeking* respectively). Utterances with no direct mapping to either coarse-grained or fine-grained tags were labeled as *Unknown* and discarded during training and testing.

## 4 Results and Discussion

In this section we present the results of our experimental study. We evaluate the performance of our model on two test sets, namely the test split of GDC and the DialogBank corpus. We provide average Micro-F1 and Macro-F1 scores for our model, with the former providing a measure of its raw accuracy and the latter providing a better metric for how well low-frequency tags are correctly classified. The DASHNet model was able to correctly annotate a large portion of the test split of the GDC data set. Moreover, it also shows promising results on the DialogBank test set, highlighting good generalisation when classifying out-of-domain data.

Table 2 shows the main results of our study. We compared our model DASHNet with HiAGM-TP (Zhou et al., 2020) and with the suite of SVM classifiers proposed by (Mezza et al., 2018). As the code for both systems is openly available[2] [3], we trained both on our GDC training data and tested on both the GDC test split and the entire DialogBank corpus. The DASHNet architecture outperforms both approaches on our two test sets.

| Model | DialogBank | | GDC | |
|---|---|---|---|---|
| | **Micro-F1** | **Macro-F1** | **Micro-F1** | **Macro-F1** |
| SVM (Mezza et al., 2018) | 55.8 | 49.2 | 78.3 | 58.9 |
| HiAGM-TP (Zhou et al., 2020) | 77.3 | 49.6 | 88.1 | 71.2 |
| **DASHNet** (our model) | **83.7** | **57.1** | **90.6** | **76.9** |

Table 2: Comparative study between our model and other models in the literature. All models were trained on the train split of GDC. The DASHNet architecture outperforms other models on all of our test sets.

| Model | Micro-F1 (DBank) | Precision (DBank) | Recall (DBank) |
|---|---|---|---|
| CRF-ASN (Chen et al., 2018) | 64.8 | 64.0 | 65.6 |
| HEC (Kumar et al., 2018) | 64.0 | 63.7 | 64.3 |
| CASA (Raheja and Tetreault, 2019) | 65.3 | 68.6 | 62.4 |
| HSLT (Wang et al., 2021) | 70.2 | 70.1 | 70.4 |
| **DASHNet** (our model) | **83.7*** | **85.7*** | **81.9*** |

Table 3: Comparative study between our model and the results reported by (Wang et al., 2021). Since the authors did not specify their train-test split or their target taxonomy, it is not possible to draw a direct comparison.

Table 3 shows a comparison between our model and the results presented by (Wang et al., 2021), who published classification results on the DialogBank corpus for their neural architecture and a number of state-of-the-art models for DA tagging that

[2] https://github.com/ColingPaper2018/DialogueAct-Tagger
[3] https://github.com/Alibaba-NLP/HiAGM

they replicated. While our model outperforms all their proposed architectures, it is worth mentioning that the authors did not specify their taxonomy of fine-grained DA tags, making it impossible to draw a direct comparison. Moreover, their test set only included an unspecified subset of four dialogues of the DialogBank, while our test set includes the entire corpus. Finally, their experiments were on in-domain test data, meaning that they trained and tested their model on different splits of the DialogBank data set, whereas our focus was on out-of-domain data and how to create a model that could generalise to an unseen conversational corpus.

### 4.1 Ablation Study

Table 4 shows the results of various ablation experiments to gain a better understanding of the extent to which each component of the DASHNet architecture impacts in-domain and out-of-domain classification accuracy.

| FEATURES | DialogBank | | GDC | |
|---|---|---|---|---|
| | **Micro-F1** | **Macro-F1** | **Micro-F1** | **Macro-F1** |
| Without contextual features | 80.0 | 50.4 | 87.3 | 72.9 |
| Without Tree-LSTM prior | 81.8 | 51.2 | 90.3 | 74.8 |
| Without POS tags | 83.5 | 54.3 | 90.5 | 76.0 |
| Without DEP tags | 83.6 | 52.1 | 90.6 | 72.8 |
| **All features** | **83.7** | **57.1** | **90.6** | **76.9** |

Table 4: Ablation study results on the DialogBank and GDC test sets.

Contextual features appear to give the biggest overall boost to the performance of the model, both on in-domain and out-of-domain data. This result is in line with other works in the field which highlight the importance of contextual information when classifying Dialogue Act tags (Mezza et al., 2018; Raheja and Tetreault, 2019). Grammatical features, namely POS and dependency tags, have a marginal impact on the Micro-F1 score on both our test sets; on the contrary, they appear to have a much higher impact on the Macro-F1 score, indicating that these features are beneficial for the classification of low-frequency DA tags. Prior information about the hierarchical relationship among DA labels appears to have a significant effect on out-of-domain DA tagging, while its influence on in-domain classification appears to be more limited. This result is compatible with our assumption that taking the hierarchical nature of the taxonomy into account helps with the generalisation of the model.

### 4.2 Training Set Variations

Table 5 shows the results of our experiments with various combinations of our dialogue corpora, in order to gain a better understanding of how each resource contributed to our final results, demonstrating the benefit of a general, cross-domain corpus. We trained the network on various subsets of GDC and tested the results on the DialogBank test set.

| **Training set** | **Micro-F1 (DialogBank)** | **Macro-F1 (DialogBank)** |
|---|---|---|
| Without SWDA | 77.3 | 48.9 |
| Without MRDA | 82.7 | 50.5 |
| Without DailyDialog | 83.6 | 53.0 |
| Without Mastodon | 79.6 | 52.8 |
| Without IAC | 83.4 | 54.9 |
| **Full GDC** | **83.7** | **57.1** |

Table 5: Data set variation experiments.

Our empirical results confirmed that each and every corpus in our collection contributed to some degree to the final accuracy of the model. SWDA and MRDA, being by far the largest resources in our aggregated training corpus, appear to have a significant impact on the model's Micro-F1 and Macro-F1 scores. The Mastodon corpus proved surprisingly impactful on the testing results given its small size; a possible explanation is that its annotation scheme was designed by taking the ISO standard into account (Cerisara et al., 2018), which makes the mapping less noisy and the resulting data points more similar to those in the DialogBank. The DailyDialog corpus and IAC appear to have a lesser effect on the classification accuracy, especially on the Micro-F1 score. However, when looking at accuracy on individual tags, their impact becomes more evident: the model trained without DailyDialog performed poorly on Action-Discussion tags when compared to the one trained on the full GDC, with a 34% increase in accuracy on the *Directive* label (from 8% to 42%) and a 15%

increase in accuracy on the *Commissive* label (from 10% to 25%). Similarly, IAC had an impact on the classification of DAs that are abundant in opinionated dialogue, with a 10% increase in the accuracy of the *Agree* label (from 25% to 35%). This is partially reflected in the Macro-F1 score decrease when training without these resources.

## 5 Conclusion

We have presented a multi-dimensional, cross-domain neural architecture for ISO-Standard Dialogue Act tagging, which leverages the hierarchical nature of the standard as well as grammatical, lexical and contextual information of the input utterances. We trained the model on a General Dialogue Corpus composed of different resources mapped to the ISO 24617-2 taxonomy, and showed how our model achieves state-of-the-art performance on out-of-domain data, which highlights its generalisation capabilities. The code and GDC data set have been released so as to help advance research on this topic. In the future, we plan to expand our work by covering more ISO-annotated corpora once they become available, as well as extending the *context encoder* module of our architecture to cover a wider context of dialogue and DA history. We also plan to experiment with different embedding mechanisms, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019).

## References


Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452.

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers, https://www.cs.rochester.edu/research/cisd/ resources/damsl/RevisedManual/, accessed on May 15, 2022.

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.

Tatiana Anikina and Ivana Kruijff-Korbayova. 2019. Dialogue Act Classification in Team Communication for Robot Assisted Disaster Response. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 399–410.

John Langshaw Austin. 1975. *How To Do Things With Words*. Oxford University Press, Oxford.

Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, Magalie Ochs, and Houda Oufaida. 2020. Two-level Classification for Dialogue Act Recognition in Task-Oriented Dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4915–4925.

Harry Bunt. 1989. Information Dialogues as Communicative Action in Relation to Information Processing and Partner Modelling. In M. Taylor, F. Néel and D. Bouwhuis, editor, *The Structure of Multimodal Dialogue*, pages 47–74. North-Holland Elsevier, Amsterdam.

Harry Bunt. 2005. A Framework for Dialogue Act Specification. *WG: Proceedings of the ACL SIGSEM Working Group on Representation of Multimodal Semantic Information.*

Harry Bunt. 2006. Dimensions in Dialogue Act Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 919–924.

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.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2548–2555.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A Semantically-Based Standard for Dialogue Annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Fang. 2016. The DialogBank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3151–3158.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue Act Annotation with the ISO 24617-2 Standard. In Deborah A. Dahl, editor, *Multimodal Interaction with W3C Standards*, pages 109–135. Springer, Cham.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T Le. 2018. Multi-task Dialog Act and Sentiment Recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics,*, pages 745–754.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence Models for Dialogue. In *Proceedings of Interspeech 2018*, pages 1011–1015.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue Act Recognition via CRF-Attentive Structured Network. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Alex Fang, J Cao, H Bunt, and Xiaoyue Liu. 2012. Applicability Verification of a New ISO Standard for Dialogue Act Annotation with the Switchboard Corpus. In *Proceedings of the EACL 2012 Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*.

Emer Gilmartin, Christian Saam, Brendan Spillane, Maria O'Reilly, Ketong Su, Arturo Calvo Devesa, Loredana Cerrato, Killian Levacher, Nick Campbell, and Vincent Wade. 2018. The ADELE Corpus of Dyadic Social Text Conversations: Dialog Act Annotation with ISO 24617-2. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4016–4022.

Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J Joachim Quantz. 1995. Dialogue Acts in VERBMOBIL. *Technical report, Universität des Saarlandes, Saarbrücken*.

Daniel Jurafsky and Elizabeth Shriberg. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13, University of Colorado at Boulder, SRI International.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue Act Sequence Labeling using Hierarchical encoder with CRF. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3440–3447.

Piroska Lendvai, Antal van den Bosch, and Emiel Krahmer. 2003. Machine Learning for Shallow Interpretation of User Utterances in Spoken Dialogue Systems. In *Proceedings of the 2003 EACL Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management*, pages 69–78.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. A Unified Syntax-aware Framework for Semantic Role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pre-Training Approach. *arXiv preprint arXiv:1907.11692*.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551.

Shachi Paul, Rahul Goel, and Dilek Hakkani-Tür. 2019. Towards Universal Dialogue Act Tagging for Task-Oriented Dialogues. *arXiv preprint arXiv:1907.03020*.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Anna Schmidt, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, et al. 2014. 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)*, pages 252–258.

Silvia Quarteroni, Alexei V Ivanov, and Giuseppe Riccardi. 2011. Simultaneous Dialog Act Segmentation and Classification from Human-Human Spoken Conversations. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5596–5599.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. Hierarchical Multi-Label Dialog Act Recognition on Spanish Data. *arXiv preprint arXiv:1907.12316*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2022. Automatic Recognition of the General-Purpose Communicative Functions defined by the ISO 24617-2 Standard for Dialog Act Annotation. *Journal of Artificial Intelligence Research*, 73:397–436.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2020. Mapping the Dialog Act Annotations of the LEGO Corpus into ISO 24617-2 Communicative Functions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 531–539.

John R Searle. 1965. What is a Speech Act? In Maurice Black, editor, *Philosophy in America*, pages 221–239. Allen and Unwin, London.

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 Institute, Berkeley, CA.

Claudia Soria and Vito Pirrelli. 2003. A Multi-Level Annotation Meta-Scheme for Dialogue Acts. In *Linguistica Computazionale, vol. XVIII-XIX*, pages 865–900.

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):339–373.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *arXiv preprint arXiv:1503.00075*.

David R Traum. 2000. 20 Questions on Dialogue Act Taxonomies. *Journal of Semantics*, 17(1):7–30.

Marilyn Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817.

Dong Wang, Ziran Li, Dongming Sheng, Hai-Tao Zheng, and Ying Shen. 2021. Balance the Labels: Hierarchical Label Structured Network for Dialogue Act Recognition. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*.

Dian Yu and Zhou Yu. 2019. MIDAS: A Dialog Act Annotation Scheme for Open Domain Human Machine Spoken Conversations. *arXiv preprint arXiv:1908.10023*.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-Aware Global Model for Hierarchical Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117.

Yucan Zhou, Qinghua Hu, Jie Liu, and Yuan Jia. 2014. Learning Conditional Random Field with Hierarchical Representations for Dialogue Act Recognition. In *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, pages 1920–1923.