Automatic Recognition of the General-Purpose Communicative Functions Defined by the ISO 24617-2 Standard for Dialog Act Annotation (Extended Abstract) *

Eugénio Ribeiro[1,2] , **Ricardo Ribeiro**[1,3] and **David Martins de Matos**[1,2]
[1]INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
{eugenio.ribeiro,ricardo.ribeiro,david.matos}@inesc-id.pt

Abstract

From the perspective of a dialog system, the identification of the intention behind the segments in a dialog is important, as it provides cues regarding the information present in the segments and how they should be interpreted. The ISO 24617-2 standard for dialog act annotation defines a hierarchically organized set of general-purpose communicative functions that correspond to different intentions that are relevant in the context of a dialog. In this paper, we explore the automatic recognition of these functions. To do so, we propose to adapt existing approaches to dialog act recognition, so that they can deal with the hierarchical classification problem. More specifically, we propose the use of an end-to-end hierarchical network with cascading outputs and maximum a posteriori path estimation to predict the communicative function at each level of the hierarchy, preserve the dependencies between the functions in the path, and decide at which level to stop. Additionally, we rely on transfer learning processes to address the data scarcity problem. Our experiments on the DialogBank show that this approach outperforms both flat and hierarchical approaches based on multiple classifiers and that each of its components plays an important role in the recognition of general-purpose communicative functions.

1 Introduction

From the perspective of a dialog system, it is important to identify the intention behind the segments in a dialog, as it provides cues regarding the information present in the segments and how they should be interpreted. According to Searle [1969], the intention behind the uttered words is revealed by the corresponding dialog acts, which he defines as the minimal units of linguistic communication. Consequently, automatic dialog act recognition is an important task in the context of Natural Language Processing (NLP), which has been widely explored over the years. The ISO 24617-2 standard for dialog act annotation was defined in an attempt to set the ground for more comparable research in the area [Bunt *et al.*, 2012]. However, annotating dialogs according to this standard is an exhausting process, especially because the annotation of each segment does not consist of a single dialog act label, which in the standard nomenclature is called a communicative function, but rather of a complex structure which includes information regarding the semantic dimension of the dialog acts and relations with other segments, among other aspects. Consequently, the amount of data annotated according to the standard is still small and the automatic recognition of the whole set of communicative functions it defines remains practically unexplored.

Among the communicative functions defined by the standard, there is a hierarchically organized set of general-purpose functions. In this paper, we explore the automatic recognition of these functions, as they are predominant and, contrarily to the dialog act labels of widely explored corpora in dialog act recognition research, they pose a hierarchical classification problem, with paths that may not end on a leaf communicative function. To approach the problem, we propose modifications to existing approaches to automatic dialog act recognition that allow them to deal with the hierarchical classification problem. These modifications focus on the ability to predict communicative functions at the multiple levels of the hierarchy, identify when the available information is not enough to predict more specific functions, and preserve the dependencies between the functions in the path. Furthermore, given the small amount of dialogs annotated according to the standard available, we rely on pre-trained dialog act recognition models by using them in transfer learning processes. This way, we can take advantage of their ability to capture generic intention information and focus on identifying that which is most relevant for recognizing the general-purpose communicative functions.

In the remainder of the paper, we start by providing an overview on related work in Section 2. In Section 3, we discuss the problem posed by the hierarchy of general-purpose communicative functions and describe our approach to their recognition. Section 4 describes our experimental setup, including the dataset, evaluation methodology, and baselines for comparison. The results of our experiments are discussed in Section 5. Finally, Section 6 summarizes the contributions of the paper and provides pointers for future work.

[*] The full article is published in the Journal of Artificial Intelligence Research 73:397–436, 2022 [Ribeiro *et al.*, 2022]

2 Related Work

Given a turn, utterance, or functional segment in a dialog, to which we will refer generically as segment, automatic dialog act recognition aims at identifying the intention behind that segment. This task has been widely explored over the years, using both classical machine learning [Král and Cerisara, 2010] and deep learning. Most approaches to the task can be summarized by the following generic four-step process: 1. Given a segment in a dialog, split it into its constituent tokens and generate adequate representations for each of them. 2. Generate a representation of the segment by combining the representations of its tokens. 3. Decorate the segment representation with context information regarding the dialog history and speaker information. 4. Use the information provided by the decorated segment representation to identify the dialog act communicated by the segment.

Tokenization is typically performed at the word level, with contextualized representations leading to the highest performance [Ribeiro *et al.*, 2019b]. However, character-level (e.g. [Ribeiro *et al.*, 2019a; Raheja and Tetreault, 2019]) and functional-level tokenizations (e.g. [Chen *et al.*, 2018]) are also able to provide relevant information. The representations of the tokens can then be combined to generate a segment representation using Recurrent Neural Networks (RNNs) (e.g. [Khanpour *et al.*, 2016]), Convolutional Neural Networks (CNNs) (e.g. [Liu *et al.*, 2017]), or Transformers (e.g. [Żelasko *et al.*, 2021]). Regarding context information, the intention behind the preceding segments is typically the most important cue for the task [Ribeiro *et al.*, 2015], but speaker and turn-taking information is also relevant [Zhao and Kawahara, 2019; Wang *et al.*, 2020].

The automatic recognition of the communicative functions defined by the ISO 24617-2 standard has only been addressed in a reduced number of studies [Cerisara *et al.*, 2018; Anikina and Kruijff-Korbayová, 2019; Blache *et al.*, 2020; Wang *et al.*, 2021]. Most of these studies focused on small subsets of communicative functions and, with the exception of our preliminary study [Ribeiro *et al.*, 2020], none of them explored the automatic recognition of the complete hierarchy of general-purpose communicative functions. Furthermore, they all addressed the task as a flat classification problem using a variation of the approaches described above.

3 General-Purpose Communicative Function Recognition

The main difference between general-purpose communicative function recognition and traditional dialog act recognition is that the former poses a hierarchical classification problem, with paths that may not end on a leaf. However, both are intention recognition problems at their core. Thus, we approach the problem by adapting existing dialog act recognition approaches to deal with hierarchical problems. This way, we build on the ability of those approaches to capture generic information regarding intention. Furthermore, this allows us to explore the use of existing pre-trained models in transfer learning processes, in an attempt to minimize the impact of the scarcity of annotated data.

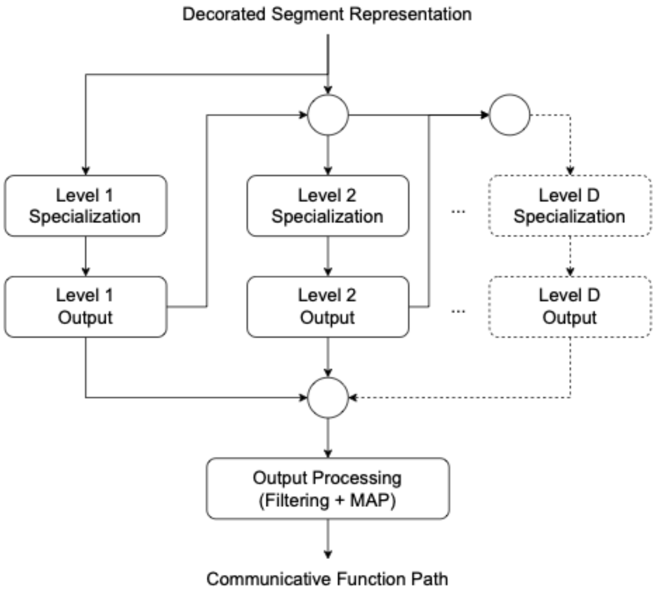

Figure 1: Our adaptation of a generic dialog act recognition approach to deal with the hierarchical problem posed by the general-purpose communicative functions. The input is a segment representation decorated with context information. The circles represent concatenation operations. *D* is the depth of the hierarchy.

As discussed in Section 2, most dialog act recognition approaches generate a segment representation decorated with context information that focuses on providing information that allows the identification of the intention behind the segment. This is also relevant for the recognition of general-purpose communicative functions. Thus, the adaptation of existing dialog act recognition approaches to the recognition of general-purpose communicative functions refers to how that decorated segment representation can be specialized to allow the identification of the hierarchically structured functions. For transfer learning purposes, in our experiments, we relied on the model that achieved the top performance on the Switchboard Dialog Act Corpus in our study on dialog act recognition [Ribeiro *et al.*, 2019b] to generate the segment representations. However, any other approach can be used.

Hierarchical classification problems are typically addressed by flattening the hierarchy or by training multiple classifiers, each specialized on a part of the hierarchy, and then using a combination of their predictions to obtain the final classification [Silla Jr. and Freitas, 2011]. On the other hand, we propose to use a single End-to-End (E2E) classifier that jointly predicts the communicative functions at each level of the hierarchy while leveraging information regarding the dependencies between them.

When dealing with the multi-class single-label classification problems posed by most dialog act annotations, neural approaches to the task include an output layer that applies the *softmax* activation function to obtain a probability distribution of the classes. The dialog act of a given segment is then predicted by selecting the class with highest probability. As shown in Figure 1, in order to consider the hierarchical structure of the general-purpose communicative functions defined by the ISO 24617-2 standard, we propose to use an output layer per level of the hierarchy instead of using a single output layer. This way, each output layer focuses on distinguishing between communicative functions at the corresponding level without having to deal with the ambiguity caused by functions that are ancestors or descendants of each other.

Additionally, we introduce a specialization layer per level, which is a fully connected layer that specializes the decorated segment representation by capturing the information that is most relevant for distinguishing the communicative functions at that level of the hierarchy. The use of specialization layers has already been proved important in our studies on multi-level dialog act recognition [Ribeiro *et al.*, 2018].

The final architectural modification refers to the use of cascading outputs, that is, the probability distribution predicted by the network at a given level is appended to the decorated segment representation before it is passed to the specialization layer of the next level. This way, the network can capture information regarding the hierarchical dependencies between communicative functions at different levels.

The general-purpose communicative functions defined by the ISO 24617-2 standard follow a strict hierarchy, the classification of a segment does not necessarily end on a leaf, and the leaves are not all at the same level. Thus, the network must be able to predict paths with variable length. To approach this problem, we add an additional label, *None*, to each level of the hierarchy, to represent that there is no communicative function attributed to the segment at that level. This way, we are able to simulate paths with fixed length, while introducing minimal impact on the network during the training phase. The drawback is that the *None* label becomes the predominant one in the deeper levels, biasing the output layers towards its prediction.

During the inference phase, the parent-child relations between the general-purpose communicative functions must be considered in order to avoid predicting an invalid path. This means that, when selecting the label at a given level, only the children of the label selected for the level above it can be considered. This restriction can be enforced using a top-down prediction approach. That is, the prediction of the path starts by selecting the communicative function with highest probability at the top level and then applies a mask on the predictions of the level below it, in order to discard the communicative functions that are not children of the selected one. This process is then repeated for each level of the hierarchy, until a leaf communicative function or the *None* label is reached. However, the performance of this approach is highly impaired when misclassifications occur in the upper levels of the hierarchy. To attenuate the impact of such misclassifications, we explore a prediction approach based on Maximum a Posteriori (MAP) estimation. That is, instead of iteratively selecting a communicative function at each level, we compute the posterior probability of all valid paths in the hierarchy and select that with highest probability. Since the classification of a segment does not necessarily end on a leaf and the leaves are not all at the same level, we also rely on the additional *None* labels during the inference phase.

4 Experimental Setup

This section describes our experimental setup, including the dataset (Section 4.1), the evaluation methodology (Section 4.2), and the baselines used for comparison (Section 4.3).

4.1 Dataset

The DialogBank [Bunt *et al.*, 2016; Bunt *et al.*, 2019] aims at collecting and providing dialogs annotated fully according to the ISO 24617-2 standard guidelines. At the time of this study, it featured annotated dialogs from four English corpora and four Dutch corpora. To avoid the issues regarding multilinguality, we focus on the 15 English dialogs, containing 2,360 annotated segments, out of which 1,118 have general-purpose communicative functions in the *Task* dimension. The distribution of general-purpose communicative functions in the DialogBank is highly unbalanced, with the most common function, *Inform*, covering 42% (62% when also considering its descendants in the hierarchy) of the segments that have a general-purpose communicative function. The distribution is also highly unbalanced across the dialogs of the different corpora that are included in the DialogBank. This reveals the heterogeneity of the DialogBank, which is representative of the different kinds of dialog that occur in human-human and human-machine interaction.

4.2 Evaluation Methodology

Although general-purpose communicative functions may occur in any of the semantic dimensions defined by the ISO 24617-2 standard, we focus on the *Task* dimension because the number of occurrences in the remaining dimensions is not representative in the DialogBank. Still, we defined two evaluation scenarios. The first considers every segment in the DialogBank, and, thus, the classifier must also decide which segments have a general-purpose communicative function. The other only considers the segments that have communicative functions in the *Task* dimension, allowing us to focus on the capabilities of our hierarchical approach.

Given the small number of dialogs in the DialogBank, we use leave-one-dialog-out cross-validation as our main evaluation approach because it maximizes the amount of gold standard data available for training.

In terms of metrics, we use the exact match ratio (MR), which considers a prediction to be correct if the whole path matches the gold standard. Additionally, we report results in terms of the hierarchical versions of precision (hP), recall (hR), and F-measure (hF) [Kiritchenko *et al.*, 2005], which also consider partial path matches.

4.3 Baselines

As baselines for comparison with our approach, we use the flat classifier approach, as well as three multiple classifier approaches to hierarchical classification: one-vs-all classifiers (OvA), one classifier per level (Lvl), and one classifier per set of siblings (Sib). In every case, the classifiers are based on the same architecture: the segment representation decorated with context information is provided to a single pair of specialization and output layers. The dimensionality of the output layer corresponds to the number of communicative functions that

	MR	**hP**	**hR**	**hF**
Flat	**74.17±.28**	79.47±.34	66.25±.23	72.26±.28
OvA	63.87±.20	**82.38±.16**	59.98±.03	69.42±.08
Lvl	69.77±.23	73.89±.19	66.58±.25	70.04±.22
Sib	69.81±.27	72.89±.47	66.84±.37	69.74±.42
E2E	72.15±.21	77.62±.23	70.99±.13	**74.16±.17**
2Step	73.30±.04	76.87±.08	**71.32±.08**	73.99±.08

Table 1: Results achieved when considering all segments.

	MR	**hP**	**hR**	**hF**
Flat	68.00±.08	85.67±.10	81.97±.12	83.77±.10
OvA	45.65±.08	**88.64±.51**	68.83±.23	77.49±.30
Lvl	56.61±.91	80.09±.66	74.43±.61	77.16±.63
Sib	55.99±.63	77.66±.55	74.35±.62	75.97±.59
E2E	**68.22±.94**	86.34±.88	**83.02±.66**	**84.65±.77**

Table 2: Results achieved when only considering the segments that have a communicative function in the *Task* dimension.

the classifier focuses on distinguishing. When using multiple classifiers, the predictions are obtained using a top-down approach. That is, starting at the root of the hierarchy, we select the communicative function with highest probability and then the process is repeated among its descendants, until a leaf communicative function or the stop case is reached. When using one-vs-all classifiers, the stop case is when none of the descendants has a probability above 50%. In the other cases, it is when the *None* label is that with highest probability.

5 Results

Table 1 shows the results of our experiments considering all the segments in the DialogBank. We can see that the approaches that rely on multiple classifiers to obtain the final predictions lead to worse performance than those that rely on a single classifier, both in terms of exact match ratio and F-measure. This suggests that, although each individual classifier is able to focus on a specific part of the hierarchy, as a whole, they are not able to capture or consider information that is important for the overall problem.

Comparing the flat approach with the E2E hierarchical one, we can see that the former achieves the highest performance in terms of exact match ratio, 74.17%, while the latter is the top performer in terms of F-measure, with 74.16%. Looking into the remaining metrics, we can see that the hierarchical approach achieves a higher recall. This means that it is able to identify communicative functions that are less prominent and/or deeper in the hierarchy. However, recall is still 6.63 percentage points below precision, which reveals a bias towards the prediction of shallower communicative functions. This can be explained by the small size of the DialogBank, which does not provide a representative coverage of the communicative functions that are deeper in the hierarchy. We were able to reduce this gap by relying on a two-step approach (2Step) that uses a binary classifier to identify the segments which have a communicative function in the *Task* dimension before applying the hierarchical approach to those segments. However, this has an impact on precision because incorrect decisions of the binary classifier cannot be corrected by the predictions on the lower levels using MAP prediction.

Using the two-step approach allows the hierarchical classifier to be trained solely on segments that have a general-purpose communicative function and, thus, be less biased towards the prediction of the *None* label. Table 2 shows the results achieved when only those segments are considered. We can see that there is a wider gap between the results in terms of exact match ratio and F-measure. One the one hand, the exact match ratio is lower because the segments without communicative function in the *Task* dimension, which are typically easy to identify, are no longer considered. On the other hand, F-measure is higher because the classifiers are less biased towards the prediction of shallower functions.

In this scenario, the E2E hierarchical approach outperforms the flat one in terms of every metric, which confirms its appropriateness. Still, the difference is only higher than one percentage point in terms of recall. However, one must consider that while the hierarchical approach covers the whole hierarchy of general-purpose communicative functions, the flat approach is trained specifically for distinguishing among the terminal communicative functions existent in the DialogBank. Thus, in a sense, the flat approach tackles a simpler problem. Considering this, the performance of the hierarchical approach, and especially its ability to identify communicative functions that are deeper in the hierarchy, suggests that it is able to capture and leverage information regarding the hierarchical dependencies as it was designed to.

Finally, our ablation studies have shown that every component of our approach plays an important role towards the recognition of general-purpose communicative functions. Still, removing the per-level specialization layers has a higher impact in performance than removing the cascading outputs and using a top-down prediction approach instead of MAP. Furthermore, given the reduced amount of dialogs in the DialogBank, transfer learning from the pre-trained dialog act recognition model plays a crucial role.

6 Conclusions

In this paper, we have explored the automatic recognition of the general-purpose communicative functions defined by the ISO 24617-2 standard for dialog act annotation. We proposed modifications to existing approaches to dialog act recognition that allow them to deal with the hierarchical classification problem posed by these communicative functions. Experiments on the DialogBank have shown that our E2E hierarchical approach outperforms a flat approach, as well as hierarchical approaches based on multiple classifiers.

In addition to the general-purpose communicative functions, the standard also defines dimension-specific communicative functions and a complete dialog act annotation includes additional information. Thus, the automatic recognition of all the relevant aspects should be addressed as future work. However, for that, additional efforts are required to increase the number of annotated dialogs that are available.

Acknowledgements

During the development of this work, Eugénio Ribeiro was supported by a PhD scholarship granted by Fundação para a Ciência e a Tecnologia (FCT), with reference SFRH/BD/148142/2019. Additionally, this work was supported by Portuguese national funds through FCT, with reference UIDB/50021/2020, and by the Recovery and Resilience Plan (RRP) and Next Generation EU European Funds, with reference C644865762-00000008 Accelerat.AI.

References

[Anikina and Kruijff-Korbayová, 2019] Tatiana Anikina and Ivana Kruijff-Korbayová. Dialogue Act Classification in Team Communication for Robot Assisted Disaster Response. In *SIGDIAL*, pages 399–410, 2019.

[Blache *et al.*, 2020] Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, Magalie Ochs, and Houda Oufaida. Two-Level Classification for Dialogue Act Recognition in Task-Oriented Dialogues. In *COLING*, pages 4915–4925, 2020.

[Bunt *et al.*, 2012] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. ISO 24617-2: A Semantically-Based Standard for Dialogue Annotation. In *LREC*, pages 430–437, 2012.

[Bunt *et al.*, 2016] Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Fang. The DialogBank. In *LREC*, pages 3151–3158, 2016.

[Bunt *et al.*, 2019] Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. The DialogBank: Dialogues with Interoperable Annotations. *Language Resources and Evaluation*, 53(2):213–249, 2019.

[Cerisara *et al.*, 2018] Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. Multi-task Dialog Act and Sentiment Recognition on Mastodon. In *COLING*, pages 745–754, 2018.

[Chen *et al.*, 2018] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. Dialogue Act Recognition via CRF-Attentive Structured Network. In *SIGIR*, pages 225–234, 2018.

[Khanpour *et al.*, 2016] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In *COLING*, pages 2012–2021, 2016.

[Kiritchenko *et al.*, 2005] Svetlana Kiritchenko, Stan Matwin, and Fazel Famili. Functional Annotation of Genes using Hierarchical Text Categorization. In *BioLINK SIG*, 2005.

[Král and Cerisara, 2010] Pavel Král and Christophe Cerisara. Dialogue Act Recognition Approaches. *Computing and Informatics*, 29(2):227–250, 2010.

[Liu *et al.*, 2017] Yang Liu, Kun Han, Zhao Tan, and Yun Lei. Using Context Information for Dialog Act Classification in DNN Framework. In *EMNLP*, pages 2160–2168, 2017.

[Raheja and Tetreault, 2019] Vipul Raheja and Joel Tetreault. Dialogue Act Classification with Context-Aware Self-Attention. In *NAACL*, pages 3727–3733, 2019.

[Ribeiro *et al.*, 2015] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. The Influence of Context on Dialogue Act Recognition. *Computing Research Repository*, arXiv:1506.00839, 2015.

[Ribeiro *et al.*, 2018] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. End-to-End Multi-Level Dialog Act Recognition. In *IberSPEECH*, pages 301–305, 2018.

[Ribeiro *et al.*, 2019a] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. A Multilingual and Multidomain Study on Dialog Act Recognition using Character-Level Tokenization. *Information*, 10(3):94, 2019.

[Ribeiro *et al.*, 2019b] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. Deep Dialog Act Recognition using Multiple Token, Segment, and Context Information Representations. *Journal of Artificial Intelligence Research*, 66:861–899, 2019.

[Ribeiro *et al.*, 2020] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. Mapping the Dialog Act Annotations of the LEGO Corpus into ISO 24617-2 Communicative Functions. In *LREC*, pages 531–539, 2020.

[Ribeiro *et al.*, 2022] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 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, 2022.

[Searle, 1969] John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

[Silla Jr. and Freitas, 2011] Carlos N. Silla Jr. and Alex A. Freitas. A Survey of Hierarchical Classification Across Different Application Domains. *Data Mining and Knowledge Discovery*, 22(1–2):31–72, 2011.

[Wang *et al.*, 2020] Dong Wang, Ziran Li, Haitao Zheng, and Ying Shen. Integrating User History into Heterogeneous Graph for Dialogue Act Recognition. In *COLING*, pages 4211–4221, 2020.

[Wang *et al.*, 2021] Dong Wang, Ziran Li, Dongming Sheng, Hai-Tao Zheng, and Ying Shen. Balance the Labels: Hierarchical Label Structured Network for Dialogue Act Recognition. In *IJCNN*, pages 1–8, 2021.

[Żelasko *et al.*, 2021] Piotr Żelasko, Raghavendra Pappagari, and Najim Dehak. What Helps Transformers Recognize Conversational Structure? Importance of Context, Punctuation, and Labels in Dialog Act Recognition. *Transactions of the Association for Computational Linguistics*, 9:1179–1195, 2021.

[Zhao and Kawahara, 2019] Tianyu Zhao and Tatsuya Kawahara. Effective Incorporation of Speaker Information in Utterance Encoding in Dialog. *Computing Research Repository*, arXiv:1907.05599, 2019.