What Helps Transformers Recognize Conversational Structure? Importance of Context, Punctuation, and Labels in Dialog Act Recognition

Piotr Żelasko[†‡], Raghavendra Pappagari[†‡], Najim Dehak[†‡]

[†]Center of Language and Speech Processing,

[‡]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA

piotr.andrzej.zelasko@gmail.com

Abstract

Dialog acts can be interpreted as the atomic units of a conversation, more fine-grained than utterances, characterized by a specific communicative function. The ability to structure a conversational transcript as a sequence of dialog acts—dialog act recognition, including the segmentation—is critical for understanding dialog. We apply two pre-trained transformer models, XLNet and Longformer, to this task in English and achieve strong results on Switchboard Dialog Act and Meeting Recorder Dialog Act corpora with dialog act segmentation error rates (DSER) of 8.4% and 14.2%. To understand the key factors affecting dialog act recognition, we perform a comparative analysis of models trained under different conditions. We find that the inclusion of a broader conversational context helps disambiguate many dialog act classes, especially those infrequent in the training data. The presence of punctuation in the transcripts has a massive effect on the models' performance, and a detailed analysis reveals specific segmentation patterns observed in its absence. Finally, we find that the label set specificity does not affect dialog act segmentation performance. These findings have significant practical implications for spoken language understanding applications that depend heavily on a good-quality segmentation being available.

1 Introduction

Human dialog is a never-ending source of diversity, abundant with exceptions and surprising ways to express one's thoughts. As a community, we have spent a massive effort in the past few decades to help the machine achieve even the slightest level of understanding of our means of communication. Remarkably, to some extent, we have succeeded. A consequence of this fact is the widespread presence of so-called *voice assistants*, that is, conversational agents of limited capabilities, which have gained much popularity in recent years.

While the main focus of modern dialog research is placed on these human–machine interactions, it is the conversation between humans that poses the greatest challenges to spoken language understanding. Consider the task of intent recognition—in a goal-oriented dialog, where the human expects their machine interlocutor to have only limited understanding capabilities, one can reasonably expect there to be a single, self-contained and straightforward utterance expressing the person's request. Siegert and Krüger (2018) show in a subjective evaluation of Alexa users that they consider such a conversation ''more difficult'' than talking to a human. With a simpler dialog structure, it is natural to approach intent recognition as a multiclass classification task, by classifying each utterance's underlying intent.

The same task of intent recognition becomes much more complex when the dialog involves two or more humans. Their conversations are riddled with various disfluencies, such as discourse markers, filled pauses, or back-channeling (Charniak and Johnson, 2001). Shalyminov et al. (2018) propose multitask training for a disfluency detection model capable of spotting hesitations, prepositional phrase restarts, clausal restarts, and corrections. Spontaneous dialogs are also characterized by much more dynamic structure than written text data. Kempson et al. (2000, 2016) show that dialog may be viewed as a sequence of incremental contributions—called *split utterances*—rather than complete sentences, and propose the *Dynamic Syntax* paradigm, claiming that standard syntactic models are insufficient to capture dialog. Another study (Purver et al., 2009) finds that up to 20% of utterances in the British National Corpus (Burnard, 2000) dialogs fit the definition of *split utterances*, with about 3% of them being cross-speaker utterance completions. Eshghi et al.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 1163–1179, 2021. https://doi.org/10.1162/tacl_a_00420
Action Editor: Claire Gardent. Submission batch: 4/2021; Revision batch: 7/2021; Published 10/2021.

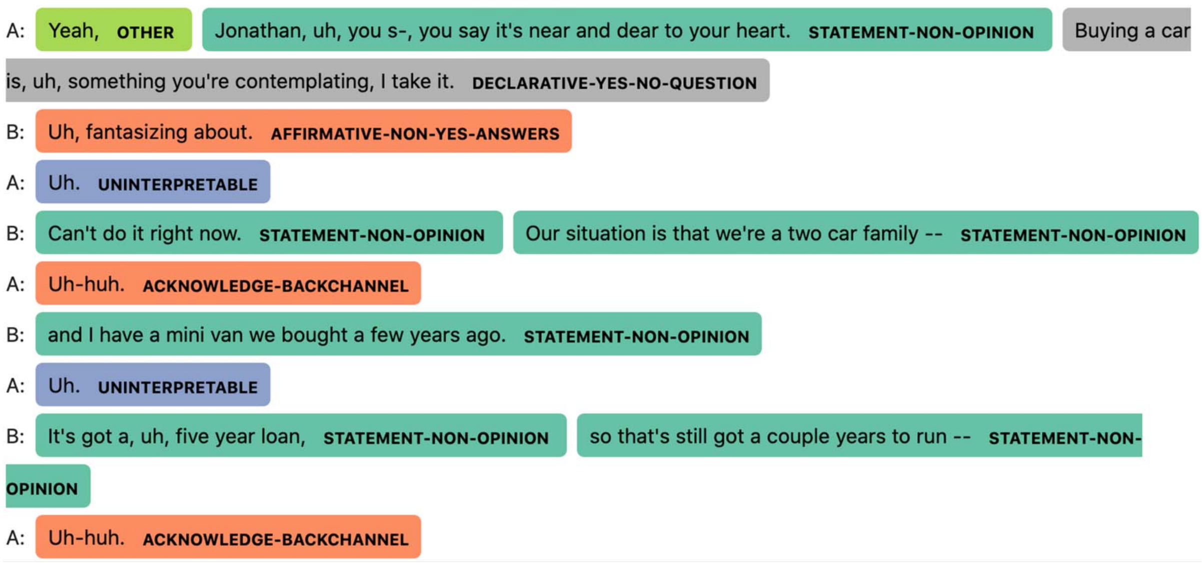

Figure 1: An illustration of dialog acts in a Switchboard conversation. Note how the speaker turns may consist of multiple dialog acts, indicating a different function for each utterance. Dialog act annotation allows us to segment the conversation into meaningful units that can be used for downstream processing in spoken language understanding (SLU) applications.

(2015) propose to view backchannels and other discourse markers as feedback in conversation that is a core component of its semantic structure, rather than a nuisance in downstream processing. This point is further argued by Purver et al. (2018), who propose incremental models for detecting miscommunication phenomena in human–human conversations. Clearly, an attempt to determine a person's intent grows beyond a turn-level classification task in such scenarios.

Dialog acts are vital to understanding the structure of the dialog. One of their modern definitions states that they are atomic units of conversation, which are more fine-grained than utterances and more specific in their function (Pareti and Lando, 2018). The part of utterance that forms a dialog act is also known as a functional segment. Recently, the definition, taxonomy, and annotation process of dialog acts has been standardized through an ISO norm (Bunt et al., 2012, 2017, 2020). Earlier studies on this topic typically used custom-tailored dialog act sets—notably, this category includes the Dialog Act Markup in Several Layers (DAMSL) scheme (Core and Allen, 1997), which was later adopted and modified to annotate the Switchboard corpus (Jurafsky et al., 1997; Stolcke et al., 2000), illustrated in Figure 1. Interestingly, dialog acts are related to the philosophy of language *speech acts* theory introduced initially by Austin (1962), in the sense that they view utterances as actions performed by the speakers.

Dialog act recognition typically entails two tasks: dialog act segmentation (DAS) and dialog act classification (DAC). In this work, we address both of them jointly and refer to their combination further as dialog act recognition. At the time of the conception of the first widely studied corpus for this task, the Switchboard Dialog Act (SWDA), DAS was considered a problem too difficult to address, and the pioneering research focused solely on the classification of dialog acts given the oracle segmentation (Stolcke et al., 2000). More recent work attempts to retrieve the segmentation through conditional random fields (CRFs) or recurrent neural networks (RNNs). However, these models still suffer from a significant margin of error, as shown by Zhao and Kawahara (2019) and later in Section 5.1. It is worth noting that in some downstream applications, the availability of high-quality segmentation is valuable regardless of any classification errors: Some examples include intent classification (Pareti and Lando, 2018), semantic clustering (Bergstrom and Karahalios, 2009), or temporal sentiment analysis (Clavel and Callejas, 2015), all of which heavily depend on the segmentation.

To the best of our knowledge, the DAS performance of transformer models (Vaswani et al., 2017) has not yet been investigated. Transformers recently demonstrated state-of-the-art performance across a range of natural language processing (NLP) tasks when combined with language model pre-training (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Beltagy et al., 2020). A major obstacle in applying transformer models to DAS is their $O(n^2)$ computational complexity with respect to the input sequence length, making it infeasible to process conversations longer than a couple of hundred tokens. Thus, there are few transformer applications to segmentation tasks—for example, Glavas and Somasundaran (2020) employed transformers for topic segmentation, but they assume that text had already been segmented and use the sentence representations instead of word representations as input to transformers.

To address the transformers' limitations, we investigate two approaches. In the first one, we use XLNet (Yang et al., 2019), a model based on the TransformerXL architecture (Dai et al., 2019), which is capable of processing the input sequence in windows while propagating the activations of the intermediate layers across as additional inputs in the following window. In the second approach, we use Longformer (Beltagy et al., 2020), which processes the whole sequence in a single pass, but for each token attends only to neighboring N other tokens, reducing the complexity to $O(mn)$, which is linear with respect to the input length.

Furthermore, we ask several questions to better understand the factors affecting dialog act recognition and design the experiments accordingly:

- *What is the significance of seeing a larger context in dialog act recognition?* Contextual dialog act models have been considered before, but they were either classification models with oracle segmentation or segmentation models that look at a limited number of past turns (see Sections 2.3 and 2.4).

- *How strongly does text formatting, that is, the presence of punctuation and capitalization, affect the segmentation quality?* This question is of significant practical importance—speech transcripts are often obtained through an automatic speech recognition system, and many of them do not offer enhanced text formatting capabilities.

- *How do the size and the specificity of the dialog act label set affect the recognition difficulty?* In some applications, the segmentation itself might be more important than having a dialog act label—for example, when clustering utterances to discover the expressions with similar meaning. Would a large, detailed dialog act label set still be beneficial for such scenarios? Are dialog act labels necessary at all, or is it sufficient to know when they begin and end?

2 Related Work

2.1 Switchboard Dialog Act

The most widely studied dialog act dataset is Switchboard (SWDA) (Jurafsky et al., 1997, 1998). It consists of telephone conversations, first manually segmented into turns and *utterances*—later formally called *functional segments* (Bunt et al., 2012), that is, the units of dialogue act annotation. Bunt et al. (2012) define them as a minimal stretch of behavior with one or more communicative functions. The total word count is about 1.4M. The conversations have 1454 words on average, and the longest one has 3122 words. The Switchboard annotators originally used the DAMSL labeling scheme (Core and Allen, 1997) with 220 dialog acts and clustered them after annotation into a reduced label set. There seems to be no consensus on the reduced label set size—some of the studies using a 42-label set are Quarteroni et al. (2011); Liu et al. (2017a); Ortega and Vu (2018); Kumar et al. (2018), others use a 43-label set (Ortega and Vu, 2017; Raheja and Tetreault, 2019; Zhao and Kawahara, 2019; Dang et al., 2020).

2.2 Meeting Recorder Dialog Act

Meeting Recorder Dialog Act (MRDA) (Shriberg et al., 2004) is a corpus of 75 meetings that took place at the International Computer Science Institute. The conversations involve more than two speakers and are significantly longer than those in SWDA. The mean word count is about 11k, and the longest dialog has 22.5k words. There are 850k words in total, making MRDA approximately half the size of SWDA. The dialog act labeling scheme is different from that in SWDA—the annotators used a 51-act set that significantly overlaps with SWDA-DAMSL (we refer to that as the *full* set). These acts were later clustered, with two granularity levels, into a *general* set of 12 acts and a *basic* set of 5 acts. The *basic* set is reduced to the following classes: *Statement*, *Question*, *Backchannel*, *Disruption*, and *Floor-Grabber*. We refer the reader to Shriberg et al. (2004) for a detailed comparison of dialog act classes between SWDA and MRDA.

2.3 Dialog Act Classification

There are two main groups of studies: The first assumes that the segmentation is known and considers dialog act recognition as a pure classification task. The original SWDA authors first take such an approach with a hidden Markov model (HMM) (Jurafsky et al., 1998). Others have introduced CRFs to solve this task (Quarteroni et al., 2011). Some authors found that considering the context explicitly in RNN models helps dialog act classification (Ortega and Vu, 2017; Liu et al., 2017a; Kumar et al., 2018; Raheja and Tetreault, 2019; Dai et al., 2020). Also, it has been shown that incorporating acoustic/prosodic features helps as well to some extent (Ortega and Vu, 2018; Si et al., 2020). Colombo et al. (2020) report the best result to date for SWDA classification—an accuracy of 85%, obtained by a sequence-to-sequence (seq2seq) GRU model with guided attention. For MRDA, the best classification accuracy is 92.2% reported by Li et al. (2019), achieved with a dual-attention hierarchical bidirectional gated recurrent unit (BiGRU) with a CRF on top. These approaches are not directly comparable with ours, as they assume an oracle segmentation of the transcript.

2.4 Dialog Act Segmentation and Recognition

More interesting in the context of our work are the studies that consider dialog act segmentation and recognition. One of the first attempts was made by Ang et al. (2005) with decision trees and HMMs for the MRDA corpus. CRF has been successfully used in this task (Quarteroni et al., 2011). The closest work to ours is by Zhao and Kawahara (2019), where a BiGRU model is used to segment and classify dialog acts in SWDA jointly. The model is considered as a sequence tagger with an optional CRF layer or in an encoder–decoder setup. It also integrates previous dialog act predictions for ten previous turns using an attention mechanism. Notably, the main differences from our setup are that Zhao and Kawahara (2019):

1. consider prediction for a single turn at a time, whereas our dialog-level contextual models process multiple turns at the same time, which allows to include both past and future context into prediction;

2. use exclusively lowercase text without punctuation, whereas we study setups both with and without the punctuation and truecasing;

3. limit the vocabulary at 10000 words, whereas we use sub-word tokenizers with no such limitation—this results in the model being able to leverage another 10000 less-frequent words in SWDA, which would have otherwise been replaced by an out-of-vocabulary symbol;

4. connect dialog act continuations (the segments labeled in SWDA with a $+$) to the previous turn when interrupted, for example, by a backchannel—we view that operation as a work-around for their models to be able to see the relevant future context, whereas our proposed models require no such pre-processing.

Finally, we provide a more detailed analysis of the effect of context on the recognition outputs; we also investigate the effect of punctuation and label set specificity, which is not discussed in that work.

2.5 The Effect of Context and Punctuation

In Liu et al. (2017b), the authors process each dialog act segment in parallel streams using a CNN and combine the sequence of sentence representations using an LSTM to exploit the context. The influence of context is explored in Bothe et al., (2018) by using an LSTM on the segment representations. Here, dialog act classification is achieved in two stages: learning segment representations and dialog act classification using an LSTM. The usage of punctuation marks as features, and other heuristics such as the number of words in the segment, *n*-grams, the dialog act of the next segment, and others, is explored in Samuel et al. (1998) and Verbree et al. (2006). However, the effect of each of these heuristics, especially punctuation marks, is not analyzed. To the best of our knowledge, there are no studies that attempt to understand the role of context, punctuation, or label set specificity on dialog act recognition in-depth.

3 Methods

3.1 Transformers

The transformer architecture is shown to produce state-of-the-art results on several NLP tasks

(Vaswani et al., 2017; Devlin et al., 2019). It consists of repeated blocks of a self-attention layer and a feed-forward layer. The self-attention layer processes the entire input sequence and learns to attend to the relevant tokens by computing the cross-token similarity in the input sequence. The similarity computation is implemented with a dot-product followed by a softmax operation. Each token's representation in the self-attention layer output is passed through a feed-forward layer before the next self-attention layer. However, as the self-attention layer processes all tokens of the input sequence simultaneously, it is invariant to the input sequence's token order. The ordering information is preserved by adding positional embeddings to the input token embeddings. Positional embeddings include one vector per token position and are learned during model training together with other model parameters.

One major limitation of transformer models is their scalability to longer inputs, as the complexity of each self-attention layer is $O(n^2)$ where n is the input sequence length. More recent work addresses this limitation in several ways: 1) propagation of context between segments of long sequence (Dai et al., 2019; Yang et al., 2019), 2) local attention (Ye et al., 2019; Beltagy et al., 2020; Wu et al., 2020; Zaheer et al., 2020), 3) sparse attention (Kitaev et al., 2020; Tay et al., 2020; Zaheer et al., 2020), and 4) efficient attention operation (Wang et al., 2020; Katharopoulos et al., 2020; Shen et al., 2021). In this work, we explore two of these models for dialog act recognition: XLNet (Yang et al., 2019) which is based on the propagation of context, and Longformer (Beltagy et al., 2020) which uses local attention.

3.2 XLNet

XLNet (Yang et al., 2019) is a transformer model trained with a masked language model criterion. It consists of 12 (*base*) or 24 (*large*) self-attention layers. It is based on TransformerXL (Dai et al., 2019), which enables it to process text sequences in windows while propagating the context in the forward direction. We leverage this property to process conversational transcripts efficiently. Furthermore, XLNet is pre-trained as an autoregressive language model that maximizes the expected likelihood over all permutations of the input sequence factorization order. It is interesting to note that this model, unlike BERT, uses relative positional encodings that do not need to be learned, making it possible to process sequences of arbitrary lengths. Even then, the quadratic computational complexity necessarily renders such processing infeasible, making windowed processing a more practical choice.

3.3 Longformer

Longformer (Beltagy et al., 2020) is based on a modification of the self-attention layer that reduces the computational complexity by limiting the context available to each input token. It splits the attention into two components—local and global. The local component is a sliding window of fixed size for each self-attention layer, dramatically reducing long sequences' computational complexity. The global component allows select tokens to attend to the entire sequence. We do not use it in this work—unlike in text classification, where [CLS] uses global attention, or question answering, where the question tokens use global attention (Beltagy et al., 2020), there are no clear candidates for it in dialog act recognition. Following Beltagy et al. (2020), we use RoBERTa (Liu et al., 2019) (BERT with carefully tuned hyperparameters) as the base model to avoid the costly pre-training process. This model's limitation is that it cannot process token sequences longer than those seen during training (4096 tokens for the pre-trained model open-sourced by Beltagy et al. [2020]). We investigate Longformer because we consider its sliding window attention mechanism as a natural extension over the XLNet's window-processing mechanism.

4 Experimental Setup

4.1 Model Training

For both transformer models, we use pre-trained sub-word tokenizers and weights, as provided by HuggingFace[1]—*allenai/longformer-base-4096* for Longformer and *xlnet-base-cased* for XLNet. These are the *base* variants with 12 self-attention layers. To adapt the models to the DAS task, we put a token classification layer on top of the transformer and train it with a per-token cross-entropy loss. We fine-tune each model on the training portion of the dataset—1003 calls for SWDA and 51 meetings for MRDA. We use the validation set (112 SWDA calls; 12 MRDA meetings) to select the best model for each variant

[1] `https://huggingface.co/`.

and the test set (19 SWDA calls; 12 MRDA meetings) for the final evaluation.

The baseline BiGRU model is trained in the same setup as described in Zhao and Kawahara (2019). For both XLNet and Longformer, we compare their performance to BiGRU by training them as turn-level models that see only a single speaker turn without additional context. In a separate experiment, to measure the effect of providing the surrounding dialog context, we train them as broad-context models processing either full transcripts (Longformer) or chunks (XLNet). All reported metrics are the mean values from three runs with different random seeds (42, 43, 44).

We train each model with a single GeForce GTX 1080 Ti GPU, which allowed us to construct batches of 6 chunks with 512 tokens each for XLNet training. The same setup might not be optimal for Longformer, as only the first 512 positional embeddings would have been fine-tuned. Therefore, we train it with 4096 token windows and an effective batch size of 6, using gradient accumulation. All models are trained for ten epochs with an Adam optimizer, a learning rate of 5e-5, and a learning schedule linearly decreasing its value towards 0. We evaluate the model on the validation set after each epoch and select the model that achieved the best F1 macro score to report the test set results.

4.2 Data Preparation

To transform the SWDA[2] and MRDA[3] conversational transcripts into model inputs, we perform several steps. First, we remove all annotator comments from the SWDA text. We evaluate each model in two variants, with/without punctuation and truecasing, to investigate how strongly it affects the performance. When punctuation and truecasing are used, they are always the ground truth. To create a single sequence out of speaker turns, we concatenate them with a unique *TURN* token in between that does not participate in loss computation but explicitly indicates that the speaker has changed.

Following Zhao and Kawahara (2019), we encode the dialog act labels using an **E** joint coding scheme. In the **E** scheme, each word comprising a dialog act is assigned a label; the **E** label indicates an end of the dialog act, and the **I** label indicates a token other than an ending. The joint coding also specializes the **E** label for each dialog act class in the label set, allowing to perform dialog act recognition. The **I** label is shared between all dialog act classes. BERT models typically use sub-word tokenization—byte-pair encoding (Gage, 1994; Sennrich et al., 2016) for Longformer and SentencePiece (Kudo and Richardson, 2018) for XLNet. When a word is split into multiple tokens, we assign the dialog act label only to the first token and discard the following tokens' predictions (i.e., they do not participate in loss computation and are ignored when reading predictions during inference).

For SWDA, we use the 42 dialog act labels (as *Abandoned-or-Turn-Exit* act is merged with *Uninterpretable*) encoded into 43 labels in total, including the **I** label. We experiment with all the label sets available in MRDA—*basic* with 5 labels, *general* with 12 labels, and *full* with 51 labels (6, 13, and 52 respectively when counting the **I** label). Unless otherwise specified, we always use the 5-label set for MRDA and 42 labels for SWDA.

Some SWDA dialog acts are extended across turns with a + label, for example, when somebody interrupted with a backchannel. We respect that by assigning an **I** label to the last token in the interrupted turn, thus creating a multiturn functional segment.

For inference, the calls are processed in sliding windows. With XLNet, we use a window size of 512 tokens without overlap. We compare the predictions with and without the context propagation across windows to understand its importance. With Longformer, we do not need to explicitly construct the windows, as each token's attention is limited to a local context of 256 neighboring tokens on each side.

4.3 Metrics

To measure the model performance, we use standard *micro* and *macro* weighted F1 metrics, as well as metrics explicitly evaluating the segmentation quality (Granell et al., 2010; Zhao and Kawahara, 2019):

- *Dialog Act Segmentation Error Rate* (DSER) measures the percentage of reference segments that were not recognized with perfect boundaries, disregarding the dialog act label.

[2]We use the SWDA distribution available here: `http://compprag.christopherpotts.net/swda.html`.

[3]We use the MRDA distribution available here: `https://github.com/NathanDuran/MRDA-Corpus`.

- *Segmentation Word Error Rate* (SegWER) is additionally weighted by the number of words in a given segment.

- *Dialog Act Error Rate* (DER) is computed similarly to DSER but also considers whether the dialog act label is correct.

- *Joint Word Error Rate* (JointWER) is a word count weighted version of DER.

Note that these metrics are strict: If a 3-word turn with a single *Statement* act is recognized as an *Acknowledgment* on the first word and *Statement* on the next two, the micro F1 score is 66.6%, the macro F1 score is 55.5%, but the error rate metrics are all at 100%.

For reference, when reading the dialog act metrics, the SWDA and MRDA test sets have, respectively, 4500 and 16702 functional segments. For reading micro and macro F1 scores, SWDA and MRDA test sets have 29.8K and 100.6K words.

5 Results

In this section, we present the results of our experimental evaluation. Each result table is first split into *lower* and *nolower* sections, which, respectively, stand for a lowercase transcript with no punctuation, and an original case transcript with punctuation symbols. For both scenarios, we always show the results on both MRDA and SWDA datasets.

5.1 Single Turn Context Models

We start our experiments by investigating how much improvement we can achieve by replacing a simple but established BiGRU baseline model with one of the transformer models. The baseline is trained in the same setup as in Zhao and Kawahara (2019).[4] To make the comparison fair, we train the XLNet and Longformer on single turn inputs so that the model does not see any dialog context. The same is true during inference. The results are shown in Table 1.

Both transformer models offer substantial improvements over the BiGRU baseline in all scenarios. In most evaluations, XLNet achieves the best results, outperforming Longformer by a small margin, compared to the improvement over BiGRU. Because these experiments do not test the model's ability to handle long-range context, these results suggest that XLNet's pre-training procedure is more suitable for dialog act recognition than that of Longformer.

5.2 Broad Context Models

In the second experiment, we investigate how long-document transformers perform in dialog act recognition. As a baseline (*Turns*), we re-use the best model from Section 5.1 (XLNet) processing dialog transcript on a turn-by-turn basis without additional context. The other proposed models process the whole transcript in sliding windows. XLNet uses a window of 512 tokens with a step size of 512 tokens. This window traversal strategy is not optimal—the tokens on the window boundaries cannot attend to other tokens close by but belonging to another window. XLNet+*prop* partially addresses this issue by propagating the intermediate activations between the windows. Longformer uses a window of 512 tokens with a step size of 1 token, which is possible thanks to its special local attention pattern. Therefore, it fully avoids XLNet's traversal strategy issue. The results are in Table 2.

All broad context models outperform the turn-level baseline across all metrics, except the turn-level SWDA *nolower* baseline in the JointWER metric. XLNet+*prop* emerges as the best model in all configurations with minor gains over XLNet. Similarly, as in Section 5.1, we observe consistent improvements in all setups when using XLNet instead of Longformer. However, we cannot conclude that XLNet uses the context more effectively, as its performance on context-less turn prediction was also better than that of Longformer. Besides the attention patterns, there are other differences between the models, such as the pretraining conditions and positional encoding schemes, which could also explain the observed results. However, it is an indication that limiting Longformer's number of positional embeddings to 4096 is not a limiting factor in its performance.

We compare the runtime of XLNet and Longformer models. Average inference time with 512 tokens window on SWDA transcripts with an eight-core Intel Core i9-9980HK CPU takes 2.8

[4] During replication, we discovered an issue in the experimental results reported in that paper—the segment insertion errors were not counted, which artificially lowered the error rates. We contacted the authors and agreed that the results we report for their model are the correct ones.

Case	Dataset	Model	micro_f1	macro_f1	DSER	SegWER	DER	JointWER
lower	MRDA	BiGRU	92.66	64.68	41.69	51.56	54.78	59.54
		Longformer	**94.02**	**70.25**	34.55	41.15	45.74	46.71
		XLNet	**94.02**	69.54	**33.62**	**40.40**	**45.62**	**46.38**
	SWDA	BiGRU	92.90	34.16	29.31	40.51	49.59	57.83
		Longformer	**94.04**	**41.15**	20.27	28.50	**40.29**	45.45
		XLNet	93.99	39.56	**19.79**	**27.12**	41.13	**45.18**
nolower	MRDA	BiGRU	96.60	79.21	18.28	22.31	27.91	25.67
		Longformer	97.08	80.80	16.19	18.05	25.34	20.26
		XLnet	**97.12**	**81.71**	**15.08**	**17.81**	**24.01**	**19.89**
	SWDA	BiGRU	94.47	38.92	14.21	22.31	37.86	44.62
		Longformer	95.35	**46.87**	11.00	16.21	32.31	35.78
		XLnet	**95.40**	46.24	**9.98**	**14.64**	**31.85**	**34.67**

Table 1: Dialog act recognition performance for BiGRU (baseline), XLNet, and Longformer models on SWDA and MRDA datasets. The models are processing each speaker turn separately, without seeing any additional context.

Case	Dataset	Model	micro_f1	macro_f1	DSER	SegWER	DER	JointWER
lower	MRDA	Turns†	94.02	69.54	33.62	40.40	45.62	46.38
		Longformer	94.65	75.30	32.78	39.70	44.11	45.17
		XLNet	94.82	75.49	**32.71**	38.74	43.78	44.21
		+*prop*	**94.89**	**75.82**	32.87	**38.32**	**43.61**	**43.76**
	SWDA	Turns†	93.99	39.56	19.79	27.12	41.13	45.18
		Longformer	95.51	53.70	18.60	25.17	38.60	45.55
		XLNet	95.49	53.48	17.74	24.24	37.99	44.88
		+*prop*	**95.57**	**54.86**	**17.48**	**24.09**	**37.51**	**44.38**
nolower	MRDA	Turns†	97.12	81.71	15.08	17.81	24.01	19.89
		Longformer	97.45	85.31	14.52	17.41	22.87	19.45
		XLNet	**97.57**	85.54	14.43	**16.59**	22.56	**18.59**
		+*prop*	97.55	**85.67**	**14.15**	16.85	**22.29**	18.92
	SWDA	Turns†	95.40	46.24	9.98	14.64	31.85	**34.67**
		Longformer	96.58	57.73	8.76	12.98	30.73	36.41
		XLNet	96.57	57.91	8.40	**12.28**	30.67	36.42
		+*prop*	**96.65**	**58.17**	**8.39**	12.34	**30.21**	35.90

Table 2: Dialog act recognition performance of large-context models—Longformer and XLNet. XLNet+*prop* means that the intermediate activations are passed between the processed segments during inference. †The best turn-level model, i.e., the XLNet, is used as a baseline (*Turns*).

seconds for Longformer and 14.7 seconds for XLNet, making Longformer about five times faster when deployed on a CPU. Figure 2 shows the time it takes for dialog act prediction on a 1750 words call *sw2229* from SWDA—for smaller windows of 32 and 64, the models take similar time to run, but as the window size increases, Longformer becomes quicker than XLNet. To summarize, Longformer might be more suitable for practical applications, even if it achieves slightly worse recognition results.

An analysis of confusion patterns in the most performant model (*nolower* XLNet+*prop*) does not reveal any new insights in SWDA compared with past works—the most confused label pair is *Statement-opinion* and *Statement-non-opinion*. For the same model in MRDA, we observe the *Question* label has the highest F-score of 98.32%,

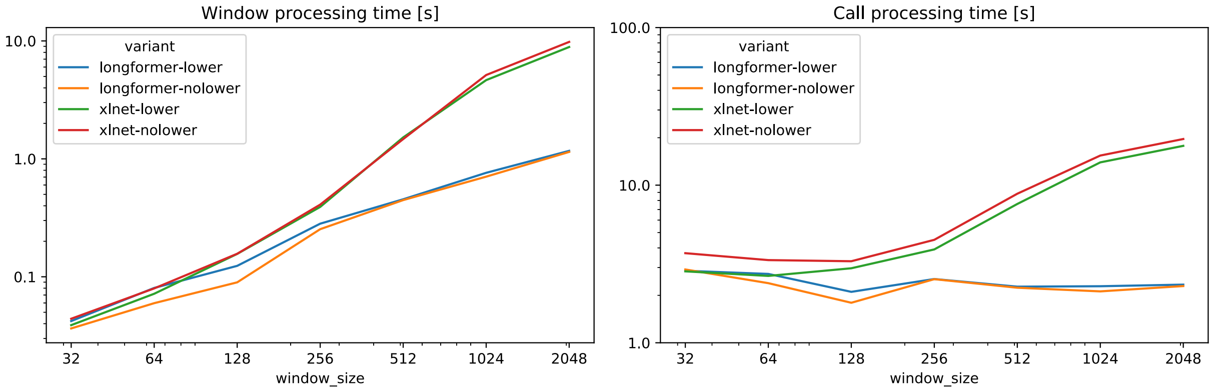

Figure 2: Prediction time for SWDA call *sw2229* by Longformer and XLNet with different window sizes. The left-side plot shows the mean time it takes to predict a single window, and the right-side plot shows the time needed to process the full dialog. Window sizes larger than 512 imply sub-windowing for Longformer, which in this experiment has learned only 512 positional embeddings.

Case	Dataset	Tagset	micro_f1	macro_f1	DSER	SegWER	DER	JointWER
lower	MRDA	51	91.90	30.94	32.93	39.15	58.62	63.90
		12	94.07	48.39	35.51	40.56	48.72	49.42
		5*	94.89	75.82	32.87	**38.32**	43.61	43.76
		1	96.74	95.23	**32.85**	38.94	–	–
	SWDA	42*	95.57	54.86	**17.48**	**24.09**	37.51	44.38
		1	98.20	97.45	17.51	24.32	–	–
nolower	MRDA	51	93.85	40.65	**13.88**	17.38	45.22	49.11
		12	96.57	64.51	14.21	17.42	27.62	26.96
		5*	97.55	85.67	14.15	16.85	22.29	18.92
		1	98.76	98.21	14.55	**16.52**	–	–
	SWDA	42*	96.65	58.17	8.39	12.34	30.21	35.90
		1	99.22	98.89	**8.37**	**12.18**	–	–

Table 3: *XLNet+prop* segmentation and recognition results for different label sets granularities; in MRDA: *full* (51), *general* (11), *basic* (5), and pure segmentation (1); in SWDA *basic* (42) and pure segmentation (1). DER and JointWER are not defined for pure segmentation. All experiments are performed using full dialog context, with identical hyperparameters, except for the output layer size. The asterisk (*) denotes the label sets typically used in other works.

followed by 94.38% for *Statements*. *Backchannels* are the most confused label, with 17% of them being classified as *Statements*, and 19% of predicted *Backchannels* being in fact *Statements*. Also, a significant portion of *Disruptions* (25%) and *Floor-grabbers* (28%) are confused with the *I* label and, respectively, 20% and 14% of them are predicted as an *I* label. This indicates that these dialog acts are the most difficult to segment correctly—which might be due to only 66.5% average inter-annotator agreement on MRDA segmentation (Shriberg et al., 2004). Lastly, 13% of predicted *Floor-grabbers* are in fact *Disruptions*.

6 Discussion

This section presents a detailed analysis of various factors affecting dialog act segmentation and recognition performance. In particular, we look into the effects of label set specificity, punctuation, and context.

6.1 The Effect of Label Set Specificity

Because MRDA provides different label set sizes, it is tempting to see how that affects the recognition performance. Furthermore, we investigate a special case where we perform pure segmentation—that is, the dialog act labels are stripped, and

Mis-segmented dialog acts	Count	DSER (turn) [%]	DSER (dialog) [%]	Abs. gain [%]
Rhetorical-Questions	12	58.3	16.7	−41.7
Other	15	53.3	20.0	−33.3
Action-directive	30	50.0	23.3	−26.7
Repeat-phrase	21	19.0	4.8	−14.3
Hedge	23	17.4	4.3	−13.0
Response-Acknowledgement	28	14.3	3.6	−10.7
Statement-non-opinion	1494	23.0	13.7	−9.3
No-answers	26	19.2	11.5	−7.7
Wh-Question	56	12.5	5.4	−7.1
Open-Question	16	6.2	0.0	−6.2
Mis-classified dialog acts	**Count**	**DER (turn) [%]**	**DER (dialog) [%]**	**Abs. gain [%]**
Yes-answers	73	100.0	17.8	−82.2
Open-Question	16	100.0	25.0	−75.0
Repeat-phrase	21	100.0	33.3	−66.7
Wh-Question	56	91.1	30.4	−60.7
Conventional-closing	84	65.5	10.7	−54.8
Response-Acknowledgement	28	89.3	35.7	−53.6
Rhetorical-Questions	12	108.3	58.3	−50.0
Collaborative-Completion	20	100.0	55.0	−45.0
Backchannel-in-question-form	21	57.1	19.0	−38.1
Summarize/reformulate	25	100.0	72.0	−28.0

Table 4: Top 10 SWDA dialog acts that benefit from dialog-level context availability in pure segmentation and dialog act recognition. The columns denoted by (turn) and (dialog) represent numbers for turn-level XLNet and dialog-level XLNet+*prop*.

there remains a single generic **E** token at the end of each segment. For SWDA, we compare 42-label set performance with pure segmentation. All experiments are performed using the XLNet+*prop* model, which was the best model in Section 5.2. The results are shown in Table 3.

We do not observe a strong effect of the label set size on segmentation performance; the pure segmentation model is practically on par with the dialog act recognition model. This is indicated by little change in DSER and SegWER metrics across the label sets in each experimental scenario. On the other hand, the label set size has a major effect on the classification performance, reflected in F1, DER, and JointWER. We offer two explanations for that. Firstly, the larger label sets have more imbalanced classes, e.g., in the 51 labels set, 43% of acts are statements, and the 18th most frequent class is already below 1% of all acts. Secondly, we suspect that the inter-annotator agreement is worse for the large label set, but the MRDA authors only reported it for the five label set (80% agreement).

6.2 The Effect of Dialog Context

To understand how the dialog context helps improve the models, we analyze the predictions of turn-level XLNet and dialog-level XLNet+*prop*. In particular, we find the subset of turns in which the turn-level model made either segmentation or classification errors, but the dialog-level model recognized everything correctly (427 turns, which is 16.3% of turns in the SWDA test set). This subset contains 752 dialog acts and suffers mostly from misclassification errors: 19.8% of these dialog acts are mis-segmented with an equal share of over- or under- segmentation, but as many as 75.8% of them have been misclassified.

We take a closer look at the differences between the two models' errors by considering the whole test set again and investigating which dialog acts benefitted the most from dialog-level context. To find them, we first have to perform segment-level alignment (since segment boundaries could be misrecognized) using the Levenshtein algorithm. For this purpose, we assume that the reference and

	Full stop	Excl. mark	Q. mark	None
Backchannel	2120 (**18.4**)	4 (50.0)	0 (0)	28 (60.7)
Disruption	115 (93.9)	2 (100.0)	6 (100.0)	2216 (**43.1**)
Floor-grabber	257 (75.5)	0 (0)	0 (0)	1152 (**49.7**)
Question	10 (20.0)	0 (0)	1231 (**8.9**)	0 (0)
Statement	9445 (14.5)	79 (**8.9**)	51 (60.8)	2 (100.0)

Table 5: Punctuation vs. dialog act counts for MRDA dataset. Percentage of errors for a given act and punctuation are shown in parentheses (the lower, the better the recognition).

predicted segments are equal when they start and end at the same words for pure segmentation and additionally check that their dialog act label is the same for recognition.

Surprisingly, we find that the strongest turn-level model (XLNet) never correctly recognized more than half of the label set (24 dialog act classes, many of which are infrequent), whereas this number significantly drops for the dialog-level model (4 classes: *Declarative-Wh-Question*, *Dispreferred-answers*, *Self-talk*, *Hold-before-answer-agreement*). The top 10 dialog acts with improved recognition performance, which occurred at least 10 times in SWDA test set, are shown in Table 4. The turn-level model lacked the necessary context to correctly classify *Yes-answers*, *Agree-Accept*, and *Response-Ackonwledgment*, mistaking them mostly for *Ackonwledge-Backchannel*. The model frequently hypothesized *Yes-No-Question* in place of *Wh-Question*. Other highly contextual dialog acts such as *Repeat-phrase*, *Rhetorical-Questions*, *Backchannel-in-question-form*, or *Summarize-reformulate* also largely improved.

In terms of segmentation performance differences, the improvements with dialog context are consistent across various kinds of dialog acts: both short (*Response-Acknowledgment*, *No-answers*) and long (*Statement-non-opinion*, *Action-directive*); questions (*Rhetorical-Questions*, *Wh-Question*, *Open-Question*) and statements.

6.3 The Effect of Punctuation – MRDA

We have previously observed from Table 3 that removing the capitalization and punctuation has a significant effect on the dialog act recognition. It suggests a strong correlation between punctuation and dialog acts. For example, a *Question* dialog act segment might often end with a question mark that could serve as a cue for the model. In this subsection, we show the correlations between dialog acts and punctuation for MRDA and SWDA datasets. Table 5 presents dialog act vs. punctuation statistics for the MRDA dataset with 5 labels. Each cell contains the frequency of a dialog act and punctuation occurring together and the percentage of our model errors in parenthesis.

We can observe that the frequency of various punctuation symbols is skewed for each dialog act. For example, segments with *Statement* and *Backchannel* dialog act labels most often contain full stop, those with *Question* dialog act label contain question mark. Similarly, *Floor-grabber* and *Disruption* labeled sentences contain no punctuation. Given that correlations between dialog acts and punctuation exist, we expect the models to leverage punctuation as a cue for prediction. Fewer errors (in bold) when punctuation is highly correlated with dialog acts confirm our hypothesis. For example, dialog act *Question* has a minimal percentage of errors when a question mark is present in the input segment. Upon further investigation, we found that the ending boundary is consistently recognized correctly when a question mark exists, and any errors that occur are at the segment's beginning. Also, the high error percentages for dialog acts *Disruption* and *Floor-grabber* could be explained due to their similar distributions of ending punctuation.

6.4 The Effect of Punctuation – SWDA

Given the large label set size of SWDA, we have no straightforward means of visualizing the correlation of punctuation and dialog acts. In order to understand the relationship between punctuation and dialog acts in SWDA, we show the top 10 most affected dialog acts in segmentation and recognition in Table 6. We observe that punctuation is key in recognizing discourse markers such as incomplete utterances, restarts, or repairs that are often labeled as *Uninterpretable*. Without punctuation,

Mis-segmented dialog acts	Count	DSER (lc) [%]	DSER (nlc) [%]	Abs. gain [%]
Rhetorical-Questions	12	58.3	16.7	−41.7
Uninterpretable	366	42.9	6.8	−36.1
Hedge	23	39.1	4.3	−34.8
Quotation	18	66.7	44.4	−22.2
Other	15	40.0	20.0	−20.0
Statement-non-opinion	1494	30.7	13.7	−17.0
Agree-Accept	213	22.1	7.0	−15.0
Statement-opinion	832	29.7	15.9	−13.8
Declarative-Yes-No-Question	38	18.4	5.3	−13.2
Open-Question	16	12.5	0.0	−12.5

Table 6: Top 10 SWDA dialog acts that benefit from punctuation and truecasing availability in pure dialog act segmentation. The columns denoted by (l) and (nl) represent numbers for dialog-level context XLNet lower and nolower models, respectively.

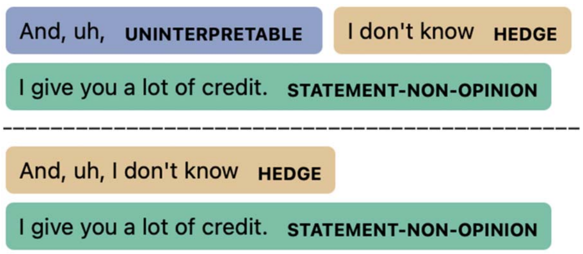

Figure 3: Top: Ground truth segmentation. Bottom: Segmentation predicted with *lower* transcripts.

these discourse markers are frequently merged into a neighboring dialog act by the model. It also partially explains the improvements in segmentation of *Statements* and some less frequent acts such as *Hedge*, since they are often found next to *Uninterpretable* (see Figure 3).

In many cases, the lack of commas takes away a cue to insert a dialog act boundary from the model. Examples are shown in Figure 4. We hypothesize that prosody or other cues found in the acoustic signal could mitigate that effect, given the usefulness of such features in dialog act classification works (Ortega and Vu, 2018; Si et al., 2020).

Another way to look at the differences in the segmentation structure is to compare the distributions of punctuation symbols found in the middle of the segments (i.e., the punctuation symbols other than the ones ending the previous and the current dialog act). We present them in Table 7. We see that the *nolower* model uses the punctuation as cues for determining segment boundary and retains a very similar distribution to the ground truth segmentation. On the other hand, the *lower*

Figure 4: Top: ground truth segmentation. Bottom: segmentation predicted with *lower* transcripts.

Segmentation	Full stop	Comma	Q. mark	Segments
ground truth	71	3637	2	4500
nolower	77	3679	2	4433
lower	155	3737	7	4323

Table 7: The number of punctuation symbols found in the middle of dialog acts, depending on the applied segmentation. *nolower* and *lower* are predicted using XLNet with dialog-level context. The presence of punctuation in *nolower* variant provides the model with the necessary cues to preserve a similar distribution to the ground truth.

model, which cannot see the punctuation, tends to under-segment the transcripts. This is consistent with our previous analyses.

7 Conclusions

We investigated how two transformer models capable of dealing with long sequences, XLNet and Longformer, can be applied to dialog act recognition. We used the well-studied SWDA and MRDA corpora and compared the performance with an established BiGRU baseline. First, we showed that the pre-trained transformers offer a substantial improvement with respect to to BiGRU when processing individual speaker turns, without any additional context. Then, we proposed adapting the transformers to consider a broader dialog context through turn concatenation with the *TURN* token, the use of joint coding, and local attention patterns or windowed processing. With this improvement, we achieved strong segmentation results on SWDA and MRDA dialog act recognition with DSER of 8.4% and 14.2% on the original transcripts and competitive results on lowercase transcripts with no punctuation (17.5% and 32.9%).

We found that XLNet was able to get the most out of the additional dialog context. We observed that the additional context is the most beneficial for segmentation while also improving the classification performance. On a practical note, Longformer allowed for approximately five times quicker inference on a modern CPU.

Across all of our experiments, it was evident that punctuation and original character cases were crucial for both segmentation and classification performance. No other factor influences the results as much—the best lowercase-transcript model (broad context XLNet+*prop*) still lags behind the simplest unmodified-transcript model (turn-context BiGRU). We analyzed the effect of punctuation and found that it is often correlated with some dialog act classes. The model leverages punctuation as a cue, especially to insert segment boundaries, but to a lesser extent also to classify dialog acts (e.g., question marks in questions).

By considering different dialog act label sets available in MRDA and a pure segmentation task, we found that XLNet's segmentation performance does not depend on the dialog act labels, further with segmentation experiments on SWDA. Regardless of the label set size (or whether the task is pure segmentation), the model performs just as well.

Finally, we found that the addition of broader context is beneficial for the model to learn rare dialog act classes—without it, more than 50% of dialog act classes were never correctly recognized even once in SWDA. With the inclusion of context, that number decreased to less than 10%.

Our findings have significant practical implications for applications that depend on text segmentation, such as the automatic discovery of intents and processes in a given domain or building graphs describing conversational flow from unstructured transcripts. We have shown that the dialog act labels do not have to be specific in order to be able to retrieve good segmentation automatically. This can significantly ease the annotation efforts, removing the need to memorize large label sets for the annotators. Furthermore, we show that the current pre-trained transformer models suffer from limitations when punctuation is not available. They tend to under-segment the text, often merging disfluencies with neighboring dialog acts. While these phenomena would likely affect, for example, systems trying to measure the semantic similarity of two segments, we expect that even the segmentation predicted on lower-case text would be useful in practical applications. It is interesting to see whether automatically retrieved punctuation can mitigate the gap between manual annotation and no punctuation; we consider this a promising candidate for future work.

To foster further research in this direction, we make our code available under the Apache 2.0 license.[5]

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1061. IEEE.

John L. Austin. 1962. *How to Do Things with Words*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150 [v1]*.

Tony Bergstrom and Karrie Karahalios. 2009. Conversation clusters: Grouping conversation topics through human-computer dialog. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2349–2352. https://doi.org/10.1145/1518701.1519060

[5] https://github.com/pzelasko/daseg/tree/version/tacl2021.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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, and Alex Chengyu Fang. 2017. Revisiting the ISO standard for dialogue act annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. 2020. The ISO standard for dialogue act annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services Oxford.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. `https://doi.org/10.3115/1073336.1073352`

Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93. `https://doi.org/10.1109/TAFFC.2015.2444846`

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *AAAI*, pages 7594–7601. `https://doi.org/10.1609/aaai.v34i05.6259`

Mark G. Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56, pages 28–35. Boston, MA.

Zhigang Dai, Jinhua Fu, Qile Zhu, Hengbin Cui, Yuan Qi, et al. 2020. Local contextual attention with hierarchical structure for dialogue act recognition. *arXiv preprint arXiv:2003.06044 [v1]*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Viet-Trung Dang, Tianyu Zhao, Sei Ueno, Hirofumi Inaguma, and Tatsuya Kawahara. 2020. End-to-end speech-to-dialog-act recognition. *Proceedings of Interspeech 2020*, pages 3910–3914.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 261–271, London, UK. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Goran Glavas and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. *ArXiv*, abs/2001.00891 [v1].

Ramón Granell, Stephen Pulman, Carlos Martínez-Hinarejos, and José Miguel Benedí. 2010. Dialogue act tagging and segmentation with a single perceptron. In *Eleventh Annual Conference of the International Speech Communication Association*.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL Labeling Project Coder's Manual*.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema DoD. 1998. *Johns Hopkins LVCSR Workshop-97 Switchboard Discourse Language Modeling Project Final Report*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–276. `https://doi.org/10.1515/tl-2016-0011`

Ruth Kempson, Wilfried Meyer-Viol, and Dov M. Gabbay. 2000. *Dynamic Syntax: The Flow of Language Understanding*. Wiley-Blackwell.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *Proceedings of International Conference on Learning Representations (ICLR)*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. `https://doi.org/10.18653/v1/D18-2012`

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017a. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D17-1231`

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017b. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178. `https://doi.org/10.18653/v1/D17-1231`

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 pretraining approach. *arXiv preprint arXiv:1907.11692 [v1]*.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 247–252. `https://doi.org/10.18653/v1/W17-5530`

Daniel Ortega and Ngoc Thang Vu. 2018. Lexico-acoustic neural-based models for dialog act classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198. IEEE. `https://doi.org/10.1109/ICASSP.2018.8461371`

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451. `https://doi.org/10.1111/tops.12324`

Matthew Purver, Christine Howes, Eleni Gregoromichelaki, and Patrick Healey. 2009. Split utterances in dialogue: A corpus study. In *Proceedings of the SIGDIAL 2009 Conference*, pages 262–271, London, UK. Association for Computational Linguistics. https://doi.org/10.3115/1708376.1708413

Silvia Quarteroni, Alexei V. Ivanov, and Giuseppe Riccardi. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5596–5599. IEEE. https://doi.org/10.1109/ICASSP.2011.5947628

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.

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*. https://doi.org/10.3115/980432.980757

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2018. Multi-task learning for domain-general spoken disfluency detection in dialogue systems. *The 22nd Workshop on the Semantics and Pragmatics of Dialogue SEMDIAL*.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. 2021. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539. https://doi.org/10.1109/WACV48630.2021.00357

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (mrda) corpus, International Computer Science Inst, Berkeley, CA. https://doi.org/10.21236/ADA460980

Y. Si, L. Wang, J. Dang, M. Wu, and A. Li. 2020. A hierarchical model for dialog act recognition considering acoustic and lexical context information. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7994–7998. https://doi.org/10.1109/ICASSP40776.2020.9053061

Ingo Siegert and Julia Krüger. 2018. How do we speak with Alexa: Subjective and objective assessments of changes in speaking style between HC and HH conversations. *Kognitive Systeme*, 2018(1).

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. https://doi.org/10.1162/089120100561737

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Daan Verbree, Rutger Rienks, and Dirk Heylen. 2006. Dialogue-act tagging using smart feature selection; results on multiple corpora. In *2006 IEEE Spoken Language Technology Workshop*, pages 70–73. IEEE. https://doi.org/10.1109/SLT.2006.326819

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768 [v3]*.

Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. Lite transformer with long-short range attention. *Proceedings of International Conference on Learning Representations (ICLR)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. 2019. BP-Transformer: Modelling long-range context via binary partitioning. *arXiv preprint arXiv: 1911.04070 [v1]*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

Tianyu Zhao and Tatsuya Kawahara. 2019. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language*, 57:108–127. `https://doi.org/10.1016/j.csl.2019.03.001`