Neural dialogue act recognition with transformer pre-training

First Author
Affiliation / Address
email@domain

Second Author
Affiliation / Address
email@domain

Dialogue acts represent the meaning of an utterance by the speech act it carries out (Austin and Urmson, 2009). Dialogue act recognition (DAR) is the task of automatically labeling utterances with tags from a dialogue act schema such as DAMSL (Core and Allen, 1997).

Dialogue acts depend on their conversational context. For example, a speaker may use *okay* to agree with their interlocutor, answer a yes-no question in the affirmative, or simply acknowledge the previous utterance, depending on what has transpired in the conversation so far. Naturally, many DAR strategies attempt to model discourse context in addition to considering the content of the utterance in question. Stolcke et al. (2000), for example, use a Hidden Markov Model to tag dialogue acts. The hidden state of neural sequence models can also be used to represent discourse context (e.g. Kalchbrenner and Blunsom, 2013; Tran et al., 2017).

It is now standard practice in NLP to initialize semantic word representations with pre-trained distributional word vectors such as word2vec (Mikolov et al., 2013). More recently, multi-layer neural language models pre-trained on massive amounts unlabeled data have been used to provide contextually sensitive word vectors and sentence-level distributional representations. One such model, BERT, uses an attention-based transformer architecture to achieve state of the art results on a variety of NLP tasks (Devlin et al., 2018).

However, given that BERT is pre-trained on book and encyclopedia data, there is no guarantee it will improve performance on dialogue-specific tasks. Adding to that uncertainty, we note that word2vec is not consistently beneficial for DAR (Cerisara et al., 2017). To assess BERT's potential for dialogue applications, we propose a series of DAR experiments with various utterance encoders, including BERT. Utterance representations are then fed to a simple RNN to predict sequences of dialogue acts (figure 1). In particular, we are interested in whether BERT can be fine-tuned to adequately represent dialogue-specific features such as discourse markers, disfluencies, and non-verbal vocalizations such as laughter.

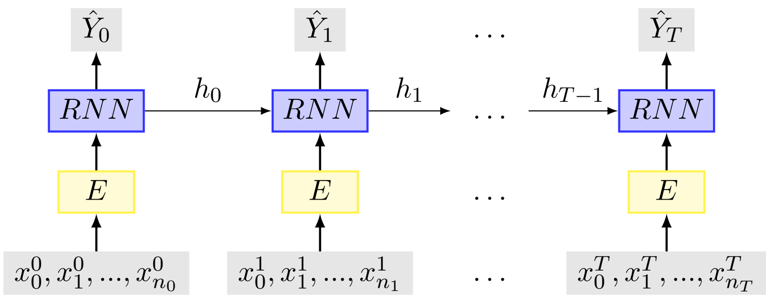

Figure 1: Simple neural DAR model. Utterance t with tokens $x_0, ..., x_{n_t}$ is encoded by E, then passed as input to the RNN. The RNN's hidden state, $ht-1$ models the discourse context at utterance t. The output layer of the RNN predicts the dialogue act, $\hat{Y}_t$.

Experiments. Our aim is to test the effectiveness of different utterance encoding strategies. We investigate the following strategies:

1. Averaging of word2vec/BERT word embedding
2. Encoding utterances with CNN (with/without word2vec/BERT initalization, with/without 'freezing' the embeddings)
3. BERT encoded sentences (with/without additional unsupervised pretraining on the full Switchboard corpus)

Pre-trained models benefit from learning on large amounts of online data, however it might be the case that for DAR, additional information only present in natural dialogue data will be useful. In SWDA disfluencies are annotated and for the datasets where no annotation for disfluencies is available, they can be predicted from recognized speech (Hough and Schlangen, 2017; Shalyminov et al., 2018). We evaluate the impact of disfluencies and non-verbal signals, such as laughter, on the performance of our models and suggest the ways to include information about disfluencies into pre-trained neural models.

References

John L. Austin and James O. Urmson. 2009. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*, 2. ed., [repr.] edition. Harvard Univ. Press.

Christophe Cerisara, Pavel Král, and Ladislav Lenc. 2017. On the effects of using word2vec representations in neural networks for dialogue act recognition. 47:175–193.

Mark G Core and James F Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

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

Julian Hough and David Schlangen. 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 326–336.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and Their Compositionality*, pages 119–126.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2018. Multi-task learning for domain-general spoken disfluency detection in dialogue systems. *arXiv preprint arXiv:1810.03352*.

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. 26(3):339–373.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. Preserving Distributional Information in Dialogue Act Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2156. Association for Computational Linguistics.