An Empirical Study of Intent Classification with Deep Classifiers

Yao P. KOUAME [*,1] and **Erwan AUCHERE** [†1]

[1]ENSAE, Palaiseau, France

March 2023

Abstract

Dialog Act Detection is currently a major research field in Natural Language Processing (NLP). Among the various methods developed to tackle this problem, the Sequence-to-Sequence architecture proposed by [1] achieves State-of-the-Art performance on the Switchboard Dialog Act (SwDA) and Meeting Record Dialog Act (MRDA) datasets. In this project, we apply this architecture to a different dataset, the Daily Dialog.

Introduction

Dialog act classification is a critical component of natural language processing, and it plays a crucial role in the functioning of virtual assistants like Siri. Essentially, dialog act classification is the process of analyzing spoken [2] or written language and identifying the underlying intent behind it. This involves breaking down each utterance or sentence into smaller units, such as individual words or phrases, and then using machine learning algorithms to classify these units based on their semantic meaning and intended function.

For virtual assistants like Siri, dialog act classification is especially important because it allows the system to accurately interpret user requests and respond appropriately. By analyzing the structure and content of user input, the system can determine whether a user is asking a question, making a statement, giving a command, or engaging in other types of communicative acts. This, in turn, allows the system to generate appropriate responses that are tailored to the user's needs and preferences [3, 4, 5, 6].

In short, dialog act classification is a crucial aspect of natural language processing, and it is essential for enabling virtual assistants like Siri to understand and respond to user requests effectively. By accurately classifying dialog acts, these systems can provide users with a more intuitive and user-friendly experience, making it easier for them to access information and complete tasks using natural language interactions.

Several works have been proposed to tackle intent classification. Among these methods is the "Linear Conditional Random Field" (CRF), which is a discriminant model (modeling is done using conditional distributions for sequence data) that models the dependence between each state (a dialogue intent) and all input sequences. We can also mention classification algorithm methods for short texts : Bag-of-Words (BoW) [7], and Continuous Bag-of-Words (CBoW) trained via a SVM model [7]. The disadvantage of these two approaches is the loss of word order and therefore the introduction of a bias in prediction [8]. To take into account the complex dependencies between words in the representation of a statement, recurrent neural networks have been introduced. More recently, Long Short-Term Memories (LSTMs) and their simplification Gated Recurrent Units (GRUs) have been used for intention classification. In our work, we approach the problem as a sequence labeling problem. We mainly rely on the

*Student engineer in DSSA: yaopacome.kouame@ensae.fr

†Student engineer in DSSA: erwan.auchere@ensae.fr

articles [9, 10, 1]. More specifically, our work is organized as follows : Section 1 is a brief literature review of the state of research in Dialog Act Detection ; Section 2 presents the architecture of the model developed in the main paper we studied ; Section 3 presents the results of our application ; and Section 4 provides a conclusion to our work and how it could be continued.

1 Related Work

Two main approaches have dominated the DA classification problem [11] : the former treats the classification of DAs as a text classification problem, where each utterance is classified in isolation ; the latter (that we will work on) approaches the problem as a sequence labeling problem.

Task as Text Classification : This method was introduced by [12]. The authors have built a vector representation for each utterance, using either a CNN or RNN, and used the preceding utterance(s) as context to classify it.

Task as a Sequence Labeling : Several authors have approached the problem in this way. Among these authors, some have used methods of CRF that only capture local dependencies as mentioned above. However, [1] consider dependencies between labels with a scope that is wider than two successive utterances. Inspired by the Neural Machine Translation (NMT), they propose Seq2seq architectures for DA analysis.

2 Dialogue Act : Sequence to sequence model

The model used by Colombo & al. (2020) for the prediction of dialogue acts is a sequence to sequence model based on a recurrent neural encoder and a recurrent neural decoder. The authors formalize the problem by considering two sets : a set of conversation denoted D which contains the set of dialogues and a set Y which contains the labels of the different dialogues. Thus, each element of D is associated to the unique value of Y. The authors consider the DA modelisation as similar to the Neural Machine Translation (NMT) task. Indeed, in NMT, the goal is to associate to any sentence $X^{l1} = (x_1^{l1} \cdots, x_{|X^{l1}|}^{l1})$ in the language $l1$ a sentence $X^{l2} = (x_1^{l2} \cdots, x_{|X^{l2}|}^{l2})$ in the language $l2$, and in DA we have $C_i = (u_1, \cdots, u_{|C_i|})$ where u_i is a sequence of words (or utterance) and we want to associate each utterance to one element of the label field $Y_i = (y_1, \cdots, y_{|C_i|})$. Thus DA can be considered as a particular case of NMT with the following specificities :

- The output vocabulary is small : whereas in NMT the output vocabulary is the actual vocabulary of the output language, in DA it consists in a few intent categories ;
- The length of the correct output sequence is known : it is the same as the length of the input sequence, which is generally not the case in NMT ;
- The input space is much larger : in NMT, the input is a sequence of words, whereas in DA it is a sequence of utterances, which are themselves sequences of words.

The task is equivalent to maximizing the likelihood of the output sequence Y_i given the input sequence C_i ($P(Y_i|C_i)$). In pratice, Colombo & al. propose a Seq2Seq architecure with a hierarchical encoder and a decoder to tackle this classification task. They introduce several types of decoders that we will present in the following sections.

2.1 Hierarchical encoder

The hierarchical encoder consists in three layers :

1. A word-level encoding that encodes each utterance u_i ($i = 1, ..., T$) of the dialogue into a vector by applying a GRU layer to the embeddings of the words composing the utterance. The representation of the utterance u_i, denoted h_i^w, is the last output of this layer ;
2. A persona-level encoding that refines the encodings of successive utterances if they are emitted by the same speaker : the whole sequence $(h_i^w)_{i=1,...,T}$ is fed to a bidirectional GRU layer whose hidden state is reset each time the speaker changes. This layer allows the representation of u_i to depend on adjacent utterances if (and only if) they share the same speaker. The outputs of this layer, a vector per utterance, are denoted h_i^p, $i = 1, ..., T$;

3. A sequence-level encoding that takes the previous representations $(h_i^p)_{i=1,...,T}$ and transform them into another sequence $(h_i^s)_{i=1,...,T}$. This layer allows the representation of u_i to depend on every other utterance in the dialogue.

The sequence $(h_i^s)_{i=1,...,T}$ is then given to the decoder.

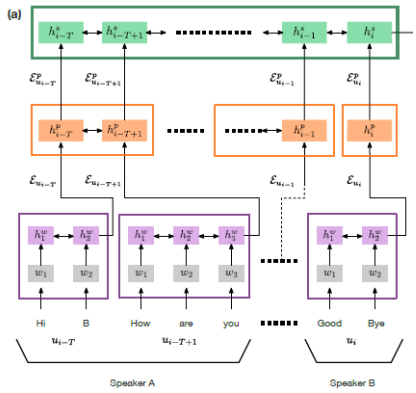

FIGURE 1 – Illustration of the structure of the hierarchical encoder (Colombo et al., 2020)

2.2 Decoders

In the decoder architecture, the authors want to force the model to learn on specific parts of the sequence each time a new word is generated and thus allow the decoder to correctly align the input sequence with the output sequence. This is done through the use of an attention mechanism : the input of the decoder's GRU layer at timestep k is defined as

$$c_k = \sum_{j=1}^{T} \alpha_{j,k} h_j^s$$

The choice of the parameters $(\alpha_{j,k})_{j=1,...,T}$, that indicates the weight that the encoder's output at timestep j has on the decoder's input at timestep k, is done through one of the three following attention mechanisms :

1. The vanilla attention :

$$\alpha_{j,k} = \text{softmax}(a(h_{k-1}^d, h_j^s))$$

where a is parametrized as a simple feed-forward neural layer and h_{k-1}^d is the output of the decoder for the previous utterance ;

2. The hard-guided attention :

$$\alpha_{j,k} = \mathbf{1}_{k=j}$$

in which we force the model to focus only on h_k^s when predicting y_k ;

3. The soft-guided attention :

$$\alpha_{j,k} = \text{softmax}(\mathbf{1}_{k=j} + a(h_{k-1}^d, h_j^s))$$

in which we push the model to focus mainly on h_k^s when predicting y_k, but allow it to consider the representation of other utterances.

Finally, the output of the decoder $((h_k^d)_{k=1,...,T})$ is the transformation by a GRU layer of the sequence of context vectors $(c_k)_{k=1,...,T}$. Each of the h_k^d is then fed into the same fully connected layer with a softmax activation to produce the output of the model.

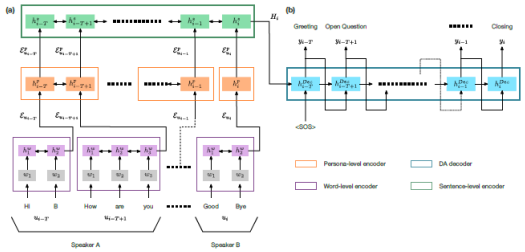

FIGURE 2 – Illustration of the structure of the model (Colombo et al., 2020)

Colombo & al. (2020) showed that hard-guided attention mechanism yielded better performances than the other two, on two different datasets : the Switchboard Dialog Act corpus (SwDA) and the Meeting Record Dialog Act (MRDA). In the next part, we will try to find the optimal attention mechanism for a different dataset, the Daily Dialog.

3 Our implementation

In this part we present our application of Seq2seq models for DA analysis. Although there are many DA dataset available [13, 14, 15, 16, 17, 18, 19, 20], we choose to focus on the Daily Dialog dataset [14]. This implementation was done in Python, with the use of the package PyTorch. The code can be found here.

3.1 The dataset

The Daily Dialog is a dataset consisting of 13118 English dialogs (11118 for training, 1000 for validation, 1000 for testing) each composed of a varying number of sentences, each sentence being associated

with one of the five labels : informative, question, directive, commissive or dummy.

Speaker	Utterance	DA label
A	Can you study with the radio on ?	question
B	No , I listen to background music .	inform
A	What is the difference ?	question
B	The radio has too many commercials .	inform
A	That's true , but then you have to buy a record player .	inform

TABLE 1 – A snippet of a conversation sample from the Daily Dialog corpus. Each utterance has a corresponding dialogue act label.

Since the model is only able to tackle dialogs of the same length, we crop the dialogs to keep only the five first utterances (we choose five as this hyperparameter is indicated to be optimal after the experiments done by [1]).

3.2 The model

The model that we used differs from the one introduced in [1] as we use a hierarchical encoder without a persona-level layer. In fact, for all dialogues in the Daily Dialog dataset, two successive utterances are always from different speakers, hence the hidden state of the persona-level would need to be reset at each timestep, leading to the intuition that this layer of the model would lose its benefit. In fact, after training both versions of the model (with and without the persona-level layer) we observed that the presence of this layer caused no improvement of the accuracy.

For the embedding layer used at the entrance of the encoder, we use pre-trained GloVe embeddings of size 50. This layer is frozen during training.

3.3 Our findings

In this study, we propose to try several decoders in the model architecture. First, we will use a decoder with a classical attention mechanism. Next, we will implement the guided attention decoders proposed in [1] : the hard-guided attention, which forces the model to focus *only* on h_t^{Enc} when predicting y_t, and the soft-guided attention, which suggests the model to focus *mainly* on h_t^{Enc} when predicting y_t but allows more flexibility. All versions of the model are trained with a hierarchical encoder without a persona-level encoder (as explained in section 3.2.) to minimize the cross-entropy loss between the predictions and the ground-truth labels. They will be compared using their accuracy on the same test set.

The following table presents the result we obtained on the Daily Dialog Act dataset (accuracy is measured on the whole generated sequence of tags, and not only on the last one) :

Decoder	Accuracy
Vanilla Attention	46.7%
Soft-Guided Attention	46.6%
Hard-Guided Attention	**46.9%**

TABLE 2 – Accuracy obtained with the different attention mechanisms

4 Conclusion

In conclusion, this work allowed us to study the richness that DA analysis can bring to Deep Learning. In the future, we would like to explore fairness of when building DA classifiers [21, 22] as these classifier are often part of systems that are deployed in open world application (*e.g.*, customer support, healthcare systems).

Références

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7594–7601, 2020.

Tanvi Dinkar*, Pierre Colombo*, Matthieu Labeau, and Chloé Clavel. The importance of fillers for text representations of speech transcripts. *EMNLP 2020*, 2020.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-driven dialog generation. *NAACL 2019*, 2019.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*, 2020.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*, 2021.

Pierre Colombo, Chouchang Yang, Giovanna Varni, and Chloé Clavel. Beam search with bidirectional strategies for neural response generation. *ICNLSP 2021*, 2021.

Sida I Wang and Christopher D Manning. Baselines and bigrams : Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 90–94, 2012.

Jetze Schuurmans and Flavius Frasincar. Intent classification for dialogue utterances. *IEEE Intelligent Systems*, 35(1) :82–88, 2019.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. Code-switched inspired losses for spoken dialog representations. In *EMNLP 2021*, 2021.

Emile Chapuis*, Pierre Colombo*, Matteo Manica, Matthieu Labeau, and Chloe Clavel. Hierarchical pretraining for sequence labelling in spoken dialog. *Finding of EMNLP 2020*, 2020.

Vipul Raheja and Joel Tetreault. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv :1904.02594*, 2019.

Ji Young Lee and Franck Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv :1603.03827*, 2016.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard : Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA, 1992. IEEE Computer Society.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog : A manually labelled multi-turn dialogue dataset, 2017.

Geoffrey Leech and Martin Weisser. Generic speech act annotation for task-oriented dialogues. 2003.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap : Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42 :335–359, 12 2008.

R. Passonneau and E. Sachar. Loqui human-human dialogue corpus (transcriptions and annotations), 2014.

Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. The hcrc map task corpus : natural dialogue for speech recognition. 01 1993.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld : A multimodal multi-party dataset for emotion recognition in conversations, 2018.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *ICML 2022*, 2022.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*, 2022.