Weak supervision for Question Type Detection with large language models

Jiří Martínek, Christophe Cerisara, Pavel Král, Ladislav Lenc, Josef Baloun

Weak supervision for Question Type Detection with large language models

Jiří Martínek[1,2], *Christophe Cerisara*[3], *Pavel Král*[1,2], *Ladislav Lenc*[1,2], *Josef Baloun*[1,2]

[1]Dept. of Computer Science & Engineering
University of West Bohemia, Plzeň, Czech Republic

[2]NTIS - New Technologies for the Information Society
University of West Bohemia, Plzeň, Czech Republic

[3]Université de Lorraine
CNRS, LORIA, F-54000, Nancy, France

{jimar,pkral,llenc,balounj}@kiv.zcu.cz, cerisara@loria.fr

Abstract

Large pre-trained language models (LLM) have shown remarkable Zero-Shot Learning performances in many Natural Language Processing tasks. However, designing effective prompts is still very difficult for some tasks, in particular for dialogue act recognition. We propose an alternative way to leverage pre-trained LLM for such tasks that replace manual prompts with simple rules, which are more intuitive and easier to design for some tasks. We demonstrate this approach on the question type recognition task, and show that our zero-shot model obtains competitive performances both with a supervised LSTM trained on the full training corpus, and another supervised model from previously published works on the MRDA corpus. We further analyze the limits of the proposed approach, which can not be applied on any task, but may advantageously complement prompt programming for specific classes.

Index Terms: Question Types Classification, Zero-shot, Dialogue Acts, BART

1. Introduction

Large pre-trained language models (LLM[1]) have shown remarkable Zero-Shot Learning (ZSL) performances in many Natural Language Processing (NLP) tasks [1], thanks to their ability to accumulate various types of information in their parameters and retrieve the correct piece of knowledge when given an input prompt that describes the target task. However, this generic prompt programming approach may be difficult to use with some tasks for which no obvious and efficient prompt exists. We focus in this work on such an NLP task: dialogue act (DA) recognition, and more specifically on question type classification, which is a subtask of DA recognition.

Dialogue act recognition is an important NLP task for automatic dialogue systems and conversational agents: it consists in tagging every spoken dialogue turn with its function in the dialogue, for example with tags such as *Request*, *Statement*, *Backchannel* and *Question* [2]. Several types of questions are often considered in DA tag sets, such as *Yes-no question*, *Wh question* and *Or question*. We have observed in preliminary experiments that directly using simple prompts with state-of-the-art LLMs, e.g.: "Is the following sentence a yes-no question ?", fails for this task. We also tried simple few-shot strategies by presenting one example of the target question type in the prompt, and searched the literature for related works that would propose adequate prompts for this task, without success. Therefore, we might probably need to rely on more advanced prompt programming strategies to solve the task of question type classification with zero of few-shot learning, which limits the interest of this paradigm for real use case application practitioners.

We thus propose an alternative approach that still leverages the ZSL capabilities of LLMs but replaces the difficult prompt designing process by intuitive and easy-to-write rules, and apply it on the question type classification task. Our proposed approach is evaluated on the meeting recorder dialogue act (MRDA) corpus [3] and its performances are compared with those of a supervised LSTM model. We further analyze the types of errors and propose future directions of research.

2. Proposed Approach

We propose to exploit a state-of-the-art pretrained LLM that is fine-tuned to perform Natural Language Inference. More specifically, we use a pretrained BART [4] model [2] finetuned on the multi-genre natural language inference (MNLI) [5] task. BART (bidirectional and auto-regressive transformer) is a denoising sequence-to-sequence autoencoder. It is trained by corrupting documents and optimizing a reconstruction loss [4]. BART has been evaluated on several tasks and can be used for a wide spectrum of downstream applications, including Question Answering (SQuAD) [6], text summarization [7, 8] and MNLI. The MNLI corpus is a collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). The premise sentences are gathered from ten different sources, including transcribed speech, fiction, and government reports [9].

BART-MNLI is a model that may be used in ZSL for various NLP tasks [10], by converting every class label into a natural language sentence that describes the label. It is then possible to infer an unknown class label according to whether the input entails the label description or not. For instance, computing the BART-MNLI entailment scores between the hypothesis "this text is positive" vs. "this text is negative" and the target paragraph as a premise gives competitive accuracy on sentiment analysis. We chose the BART-MNLI model because related works have previously shown the dependency between the sentiment analysis and DA recognition tasks [11], and we

[1]We call LLM any large pretrained NLP embeddings model, even when it is not strictly a language model that predicts the next word.

[2]https://huggingface.co/facebook/bart-large-mnli

thus expected this model to also contain relevant information for DA recognition. However, as mentioned before, it is much more difficult to devise working hypothesis for question type classification, and we therefore rather propose to encode the following intuitive "rules" to leverage the ZSL capabilities of BART-MNLI:

- R1: Yes-no questions are commonly followed by the answer "yes" or "no".
- R2: Or-Clause questions usually start with "or". Note that this rule is directly derived from the target corpus annotation guide.
- R3: In Or-questions, such as "do you prefer A or B ?", A and B often represent competing alternatives.
- R4: Wh-questions usually start with a Wh-word.

Our claim is that the simple and intuitive R1-4 rules may advantageously replace the difficult-to-design prompts to achieve ZSL for question type recognition. However, the proposed approach also has its limits, as there exists no such intuitive rule to characterize Open-ended and Rhetorical questions. This fact is supported by the target corpus annotation manual, which states that *Open-ended* questions might be also *Wh-questions* or *Yes-questions* under a condition when a specific answer is not seek, and *Rhetorical* questions are defined as question where no answer is expected. We thus let for a future work the challenge of handling both types of questions, which may require yet another ZSL paradigm beyond prompt-programming and simple rules writing given the current status of LLMs.

Our global recognition approach simply consists in applying each rule R1, R2, R3 and R4 one after the other in this order. Of course, our main contribution is not in the design of these four rules, which are excessively simple and can be written in a few minutes by any application developer, but in the proposal to exploit a powerful LLM in a zero-shot way in order to make these rules much easier to implement than with standard programming languages. In other words, thanks to their inherent semantics capabilities, LLMs enable application developers to express more abstract and high-level rules than with traditional rule-based methods. More specifically, we propose to exploit the BART-MNLI model in two ways:

1. **Next word prediction**
   The model generates the most likely following word. This generation mode is used in R1: we let the BART model generates the most probable word candidates that may follow our target question: If either *yes*, *no* or *yeah* are present among the predicted word candidates, then we label this question with the *Yes-no question* tag.
2. **Entailment score prediction**
   This configuration predicts an entailment score based on the input pair of sequences (premise and hypothesis). The model takes the input and provides an output vector with 3 scores: **contradiction**, **neutral** and **entailment**. This entailment mode is used in R3: If "or" occurs after the beginning of an utterance, we split the utterance in two parts according to the position of "or" and check whether both parts are in contradiction. For this, we use the BART-MNLI model as illustrated in Figure 1. The first part of the utterance is considered as the premise and the second part is the hypothesis. The BART-MNLI model then predicts the scores for contradiction, neutrality and entailment (CNE scores) and if the contradiction score is high enough (softmaxed score above 75%), the utterance is classified as *Or question*. The global process and the 75% contradiction threshold were designed based on prior knowledge and a few manual tests on utterances from the validation corpus.

3. Related Work

LLMs have the nice property to be able to solve various NLP tasks in a Zero-Shot way, i.e., without fine-tuning them on the target task. This is classically achieved by prefixing the input with well-designed prompts that contextualize the LLM towards the target task [1]. Advanced prompting strategies include decomposing the task-specific reasoning into several steps with *chain of thoughts* [12] and training soft prompts [13]. However this strategy seems to fail for some tasks, such as dialogue act recognition, mainly because of the difficulty to design relevant prompts. We thus propose a complementary approach that exploits frozen LLMs into programming scripts for such tasks.

The performance of supervised dialogue act recognition has significantly improved with the development of pre-trained language models and transformers. Colombo et al. [14] investigated the usage of seq2seq deep models inspired by neural machine translation with the attention mechanism. The experiments were conducted on MRDA and Switchboard (SwDA) [15] corpora with excellent results. Raheja and Tetreault [16] proposed a DA recognition model where the key components are context-aware self-attention and bidirectional GRU [17]. The input features were created by a combination of pre-trained ELMo [18] and Glove [19] word embeddings with good results on the MRDA and SwDA corpora. The more specific question type recognition task is often carried out on the TREC corpus, for instance in [20, 21]. However, the question types in TREC focus on whether a question should be answered with a person name, or a location, etc., while we focus in this work on question types that are more related to dialogue acts. In this context, a reference work has been done by Margolis and Ostendorf [22], who investigated supervised question detection on MRDA using a combination of lexical and prosodic features.

4. Application on the MRDA Corpus

The Meeting Recorder Dialogue Act (MRDA) corpus contains three levels of annotations: *basic label*, *general label* and *specific label*. The basic level of DA annotation includes five main categories, namely: *Statement*, *BackChannel*, *Disruption*, *FloorGrabber* and *Question*. In our task, we use general labels (12 labels in total). Additional information about disruption forms (indecipherable, abandoned, or interrupted) are also given if necessary. More details about the dataset and its taxonomy can be found in *Meeting recorder project: Dialog act labeling guide* [23]. The six types of questions in MRDA are shown in Table 1, along with the corresponding labels and number of sentences.

Question Type	Label	Counts
Yes-no Question	qy	804 (67.4 %)
Wh-Question	qw	259 (21.7 %)
Or-Clause	qrr	24 (2.0 %)
Or-Question	qr	28 (2.3 %)
Open-ended Question	qo	27 (2.3 %)
Rhetorical Question	qh	51 (4.3 %)

Table 1: *Question types information in MRDA test dataset*

These six types of questions are also explored by Margolis and Ostendorf in their work [22]. Questions constitute $\approx 6.5\%$ of all utterances in the MRDA corpus and the most frequent question type is *Yes-no question*.

4.1. Question Detection

Our first step consists in filtering out non-question turns from the MRDA corpus. In this corpus, every question ends with a question mark; so we first simply filter out every dialogue turn that does not end with a question mark. Then, based on the observations from [22] we further exclude the following types of sentences, which are not considered as real questions:

- Hold Before Answer/Agreement $\rightarrow$ *h*;
- Floor Holder $\rightarrow$ *fh*;
- Floor Grabber $\rightarrow$ *fg*.

This filtering pre-processing is applied similarly to the three MRDA sub-corpora (*train*, *test*, *val*) before passing them to the Question Type Classification stage.

4.2. Question Type Classification

We leverage BART-MNLI to detect whether the target sentence is a *Yes-no question*. Note that we do not access the actual following answer, but we just ask BART-MNLI whether this sentence may be followed by a yes-no answer. So writing a simple and intuitive rule such as R1 is only possible thanks to the rich information present in BART-MNLI. For R2, we search for the key word "or". According to the MRDA manual, the class *Or-Clause* is determined by the fact that it follows *Yes-no questions* and always starts with "or" (e.g. "or Saturday?", "or something?"). So a simple pattern matching rule detects the Or-Clauses. For R3, we again rely on the semantics information encoded in BART-MNLI to assess whether the two parts of the question separated by "or" are indeed alternatives.

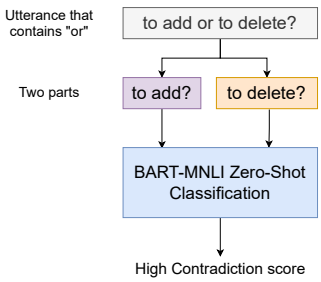

Figure 1: *Example of an Or-question*

The last remaining category is the *Wh-question*. We simply search for one of the *Wh* words[3]. If at least one of them is found at the beginning of an utterance (first three words), we predict the *Wh question* label.

4.3. Supervised LSTM-based Model

Our baseline is a supervised model trained on features computed by the same BART-MNLI model as described above. This allows us to compare two types of transfer learning from BART-MNLI: either as a ZSL model or with supervised fine-tuning. Every question is tokenized and processed by the BART-MNLI model, without any classification nor generation head, resulting into feature vectors with 1024 dimensions. These vectors are then fed into two uni-directional LSTMs layers. A softmax layer outputs the most probable question type category.

Note that the BART-MNLI feature vectors are contextual, and so, it is not really required to use a recurrent model: it is also possible to use a single feature vector to represent the whole sentence, or to pool all feature vectors by, e.g., averaging. Our preliminary experiments suggested that the LSTM was slightly better, and this is why we have chosen it next. For comparison, the baseline performance obtained when always answering the most frequent class, which is *Yes-no Question*, is also reported in our experiments (Majority Classifier).

The hidden state dimension is set to 500 in both LSTM layers. The model is trained with the Adam optimizer and the learning rate is set to 0.001. 20 training epochs are realized. These hyper-parameters have been tuned on the validation data.

5. Experiments

We present and discuss next the overall performance as well as the accuracy, precision, recall and F1-score per question type.

5.1. 4-class classification results

	Maj. Class	**LSTM**	**Zero-shot**
Precision (macro)	12.0	44.4	**56.8**
Recall (macro)	16.7	40.8	**54.1**
F1 (macro)	13.9	42.2	**54.9**
Accuracy	67.4	**89.1**	88.8

Table 2: *Performances of the proposed model and baselines on MRDA-test (4 class) [in %]*

We can see in Table 2 that our proposed ZSL approach outperforms the majority-class baseline by a large margin, and compares favorably with our supervised model, which has been trained on the full labeled training corpus. The difference between the accuracy and F1 metrics is due to the fact that the test labels are largely unbalanced. This result confirms that the proposed ZSL approach is adapted to the question type recognition task, as it successfully exploits relevant information stored in the large pretrained language model.

	qy	qw	qrr	qr
Majority class baseline				
Precision	67.4	0.0	0.0	0.0
Recall	100.0	0.0	0.0	0.0
F1	80.5	0.0	0.0	0.0
LSTM baseline				
Precision	94.5	79.8	75.0	28.3
Recall	90.3	91.5	87.5	53.6
F1	92.4	**85.3**	80.8	37.0
Proposed model				
Precision	88.6	93.2	88.5	70.4
Recall	97.6	63.3	95.8	67.9
F1	**92.9**	75.4	**92.0**	**69.1**

Table 3: *Comparison of the models per class [in %]*

Table 3 further shows that the R1 rule performs particularly well, with high recall and F1 metrics for detecting yes-no ques-

[3] how, why, whose, whom, who, which, where, when, what

tions; F1 is also comparable between the ZSL and supervised models. This confirms that BART-MNLI is a model that can correctly predict when yes-no answers are expected. Although it performs worse in absolute value when asked to detect contradictions in Or-Questions (qr), it still outperforms the supervised LSTM on the R3 rule.

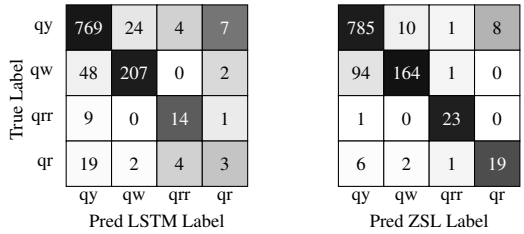

Figure 2: *Confusion matrices (4-class)*

	qy	qw	qrr	qr	qo	qh
Majority class baseline						
Precision	67.4	0.0	0.0	0.0	0.0	0.0
Recall	100.0	0.0	0.0	0.0	0.0	0.0
F1	80.5	0.0	0.0	0.0	0.0	0.0
LSTM baseline						
Precision	94.5	79.8	75.0	28.3	47.4	21.4
Recall	90.3	91.5	87.5	53.6	33.3	11.8
Selectivity	89.2	93.6	99.4	96.7	99.1	98.1
F1	92.4	**85.3**	80.8	37.0	**39.1**	**15.2**
Proposed model						
F1	**93.6**	82.2	**85.2**	**50.0**	0.0	0.0
Precision	90.7	77.5	76.7	83.3	0.0	0.0
Recall	96.8	87.6	95.8	35.7	0.0	0.0
Selectivity	79.4	92.9	99.4	99.8	100	100
Informedness	**76.2**	**80.6**	**95.2**	35.5	0	0
Margolis & Ostendorf [22]						
Informedness	69.4	76.3	83.3	**80.2**	**75.3**	**74.0**
Recall	86.1	93.0	100	96.9	92.0	90.7
Selectivity	83.3	83.3	83.3	83.3	83.3	83.3

Table 5: *Comparison of the models per class [in %]*

5.2. Extension to the 6 question types

Our proposed ZSL approach is not designed to support all 6 question types in MRDA. However, for fair comparison, we also report next the performances of our model when the Open-ended and Rhetorical questions are included. Our proposed approach can never predict both classes.

Table 4 shows the averaged performances of all approaches on the MRDA test data (all 6 classes included).

	Maj. Class	**LSTM**	**Zero-shot**
Precision (macro)	11.2	**57.7**	54.7
Recall (macro)	16.7	**61.3**	52.7
F1 (macro)	13.4	**58.3**	51.8
Accuracy	67.4	85.0	**93.1**

Table 4: *Performances on MRDA-test (6 classes) [in %]*

Conversely to Table 2, our proposed approach is now worse than the supervised LSTM on the detection metrics, which results from the fact that, by design, it does not detect any sample of the two additional classes. In terms of classification accuracy, it was comparable with the LSTM in Table 2, while it is now significantly better, which is due to the fact that the two new question types occur rarely in the corpus; when the LSTM predicts them, its classification error rate is larger for them than on the less difficult 4 previous classes, as can be seen by comparing the confusion matrices in Figures 3 and 2.

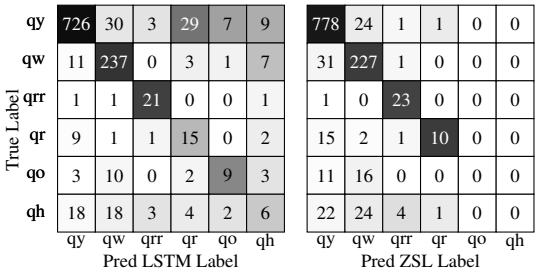

Figure 3: *Confusion matrices (6-class)*

Table 5 shows the performances per class, and further compares them with the only work we have found that reports results on the same subtask [22]: we used Informedness [24] for the latter comparison, because [22] only reports recall for a constant selectivity, so we can not compute more common global metrics from these published figures. We can see that the proposed zero-shot approach gives comparable F1 results than the supervised LSTM, and comparable Informedness results than the supervised model in [22] except for qr.

6. Conclusions

Exploiting large pretrained language models for zero-shot learning has been successfully proposed for many NLP tasks before, as listed for instance in [1]. However, the zero-shot learning paradigm is difficult to apply in some tasks, for which prompts are hard to design or fail to characterise precisely enough the target task. This is the case for dialogue act recognition, and more specifically the question type classification task. Therefore, we propose another ZSL paradigm that does not rely on prompts but on simple and intuitive rules, which are easier to design. We validate this approach with four simple rules to detect four common question types in the MRDA corpus. Two of these four rules exploit two different ZSL capabilities of the BART-MNLI model: generating the following word, and detecting contradiction between two clauses. We show that this approach, which does not involve training any supervised model, gives competitive performances with a supervised LSTM trained on the full MRDA training corpus. However, just like with prompt programming, the proposed ZSL approach is limited by its fundamental design assumption, i.e., that there exists some intuitive and simple rule that may benefit from the ZSL capabilities of LLMs to characterize the target class. Hence, we could not recognize two question types: open-ended and rhetorical questions. Solving this challenge may either require inventing a new ZSL paradigm, or waiting for LLMs to improve and become powerful enough to be able to directly answer such a prompt. In the meantime, the proposed ZSL paradigm described in this work may help to exploit LLMs in new unsupervised tasks that they could not solve so far with the standard prompt programming approach.

7. Acknowledgements

This work has been partly funded by the Lorraine Université d'Excellence project OLKi and by Grant SGS-2022-016 Advanced methods of data processing and analysis.

8. References

[1] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, "Multitask prompted training enables zero-shot task generalization," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=9Vrb9D0WI4

[2] C. Cerisara, P. Kral, and L. Lenc, "On the effects of using word2vec representations in neural networks for dialogue act recognition," *Computer Speech and Language*, vol. 47, pp. 175–193, 2018.

[3] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[5] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122. [Online]. Available: https://aclanthology.org/N18-1101

[6] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[7] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1797–1807. [Online]. Available: https://aclanthology.org/D18-1206

[8] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *Advances in neural information processing systems*, vol. 28, pp. 1693–1701, 2015.

[9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[10] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, "Entailment as few-shot learner," *preprint arXiv:2104.14690*, 2021.

[11] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. Le, "Multi-task dialog act and sentiment recognition on mastodon," in *Proc. COLING*, 2018.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *preprint arXiv:2201.11903v4*, Jan 2022. [Online]. Available: http://arxiv.org/abs/2201.11903v4

[13] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243

[14] P. Colombo, E. Chapuis, M. Manica, E. Vignon, G. Varni, and C. Clavel, "Guiding attention in sequence-to-sequence models for dialogue act prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7594–7601.

[15] J. J. Godfrey, E. C. Holliman, and J. 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*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517–520.

[16] V. Raheja and J. Tetreault, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3727–3733. [Online]. Available: https://aclanthology.org/N19-1373

[17] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[18] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] H. Tayyar Madabushi and M. Lee, "High accuracy rule-based question classification using question syntax and semantics," in *Proc. COLING*, Osaka, Japan, Dec. 2016, pp. 1220–1230.

[21] E. Cortes, V. Woloszyn, A. Binder, T. Himmelsbach, D. Barone, and S. Möller, "An empirical comparison of question classification methods for question answering systems," in *Proc. LREC*, Marseille, France, May 2020, pp. 5408–5416.

[22] A. Margolis and M. Ostendorf, "Question detection in spoken conversations using textual conversations," in *Proceedings of the ACL*, 2011, pp. 118–124.

[23] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting recorder project: Dialog act labeling guide," Intl Computer Science Inst Berkely, CA, USA, Tech. Rep., 2004.

[24] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, 2011.