PALI at SemEval-2021 Task 2: Fine-Tune XLM-RoBERTa for Word in Context Disambiguation

Shuyi Xie, Jian Ma, Haiqin Yang[§], Lianxin Jiang, Yang Mo, and Jianping Shen
Ping An Life Insurance, Ltd.
Shenzhen, Guangdong province, China
{XIESHUYI542, MAJIAN446, JIANGLIANXIN769, MOYANG853, SHENJIANPING324}@pingan.com.cn
[§] the corresponding author, email: hqyang@ieee.org

Abstract

This paper presents the PALI team's winning system for SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation. We fine-tune XLM-RoBERTa model to solve the task of word in context disambiguation, i.e., to determine whether the target word in the two contexts contains the same meaning or not. In implementation, we first specifically design an input tag to emphasize the target word in the contexts. Second, we construct a new vector on the fine-tuned embeddings from XLM-RoBERTa and feed it to a fully-connected network to output the probability of whether the target word in the context has the same meaning or not. The new vector is attained by concatenating the embedding of the [CLS] token and the embeddings of the target word in the contexts. In training, we explore several tricks, such as the Ranger optimizer, data augmentation, and adversarial training, to improve the model prediction. Consequently, we attain the first place in all four cross-lingual tasks.

1 Introduction

This year, the SemEval-2021 task 2, multilingual and cross-lingual word-in-context (WiC) disambiguation (Martelli et al., 2021), defines the task of identifying the polysemous nature of words without relying on a fixed sense inventory in a multilingual and cross-lingual setting. The task aims to perform a binary classification task to determine whether the target word contains the same meaning or not in two given contexts under both the same language (multilingual) setting and the different languages (cross-lingual) setting. In the multilingual setting, the tasks consist of English-English (En-En), Arabic-Arabic (Ar-Ar), French-French (Fr-Fr), Chinese-Chinese (Zh-Zh) and Russian-Russian (Ru-Ru) while in the cross-lingual setting, the tasks consist of English-Chinese (En-Zh), English-French (En-Fr), English-Russian (En-Ru), and English-Arabic (En-Ar).

The tasks contain the following challenges:

- The same word may deliver different meanings in different context (Lei et al., 2021).
- The training data is scarce. For example, in the multilingual tasks, there is only training data in En-En, while in the cross-lingual tasks, there is no training data.

To overcome these challenges, we explore the uniqueness of the tasks and implement several key technologies:

- First, we follow (Botha et al., 2020) to specially design an input tag for the multilingual pre-training XLM-RoBERTa model to emphasize the target word in the contexts. That is, the target word is encompassed by the special symbols of <t> and </t>. Meanwhile, the given two contexts are concatenated by the <SEP> token.
- Second, we apply data augmentation and add external data from WordNet to enrich the training data. It is noted that we only expand the data in the task of En-En and do not consider other techniques, e.g., back-translation, for the cross-lingual tasks. Adversarial training is also applied to learn more robust embeddings for target words. The Ranger optimizer with the look-ahead mechanism in the AdamW optimizer is adopted to speed up the convergence of training.
- Finally, we construct a new vector on the funetuned embeddings, i.e., concatenating the embeding of the [CLS] token and the learned embeddings of the target words' in both contexts. The new vector is then fed into a fully-connected network to produce the binary classification prediction. Cross-validation and model ensemble are also applied to attain a robust output.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 713–718
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

The rest of this paper is organized as follows: In Sec. 2, we briefly introduce related work. In Sec. 3, we detail our proposed system. In Sec. 4, we present the experimental setup, procedure, and the results. Finally, we conclude our work in Sec. 5.

2 Related Work

The SemEval-2021 task 2 aims to handling the tasks of multilingual and cross-lingual word-in-context disambiguation (Martelli et al., 2021), i.e., to determine whether the target word contains the same meaning in both given contexts. In the following, we will elaborate several related work.

Some recent effort, e.g., (Pilehvar and Camacho-Collados, 2018), has been conducted to curate and release datasets to solve the task of WiC disambiguation. Though it can be narrowed down to binary classification, some techniques have to be implemented to enhance the model performance. For example, the trick of input highlighting mechanism (Botha et al., 2020) can be facilitated to promote the importance of the target word. The idea of unifying entity linking and word sense disambiguation (Moro et al., 2014) can be borrowed to solve the task. The idea of freezing the trained model for other languages (Artetxe et al., 2020) can be explored to relieve the issue of no training data in the cross-lingual tasks.

Recently, due to the superior performance in tackling NLP tasks (Yang et al., 2021; Yang and Shen, 2021; Wang et al., 2021), pre-trained language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), start to dominate the way of word representations than static word embedding methods, e.g., Word2Vec (Mikolov et al., 2013) and FastText (Joulin et al., 2017). Especially, the XLM-RoBERTa (Conneau et al., 2020) model is a newly released large cross-lingual language model based on RoBERTa and is trained on 2.5TB filtered CommonCrawl data in 100 languages. Different from other XLM models, XLM-RoBERTa does not require the language token to understand which language is used and can determine the correct language from the input ids. It is a powerful tool for understanding multilingual languages and is very helpful for solving the WiC disambiguation task under the cross-lingual setting. Hence, we choose XLM-RoBERTa in our system.

A critical issue of the task is lack of training data. Though existing methods, e.g., lexical substitution (Zhang et al., 2015), back translation (Xie et al., 2020), and data augmentation (Fadaee et al., 2017), can be applied to enrich the data, we mainly explore the usage of WordNet (Fellbaum, 1998) and the technique of pseudo labelling (Wu and Prasad, 2018) because WordNet contains rich synonyms while pseudo labelling is effective to utilize the abundant unlabeled data via their pseudo labels.

Adversarial training (Tramèr et al., 2018) is an effective method to regularize parameters by introducing noise and to improve model robustness and generalization. We also explore its possibility in fine-tuning XLM-RoBERTa to increase the robustness of the learned the word embeddings.

3 Overview

In the following, we present the task definition, data preprocessing, and our proposed system design.

3.1 Task Defintion

The task of WiC disambiguation is framed by a binary classification task. Each instance in WiC has a target word w, whose part-of-speech is in $\{$NOUN, VERB, ADJ, ADV$\}$, with two given contexts, c_1 and c_2. Each of these contexts triggers a specific meaning of w. The task is to identify if the occurrences of w in c_1 and c_2 correspond to the same meaning or not. Figure 1 illustrates an example from the dataset.

```
{
"target word":  "play",
"sentence1":  "In that
   context of coordination
   and integration, Bolivia
   holds a key play in any
   process of infrastructure
   development.",
"sentence2":  "In schools,
   when water is needed, it
   is girls who are sent to
   fetch it, taking time away
   from their studies and
   play."
}
```

Figure 1: An example from the WiC disamguation task.

3.2 Data Preprocessing

The training dataset consists of two files in the JSON format: the .data file and the .gold file. The .data file contains the following information:

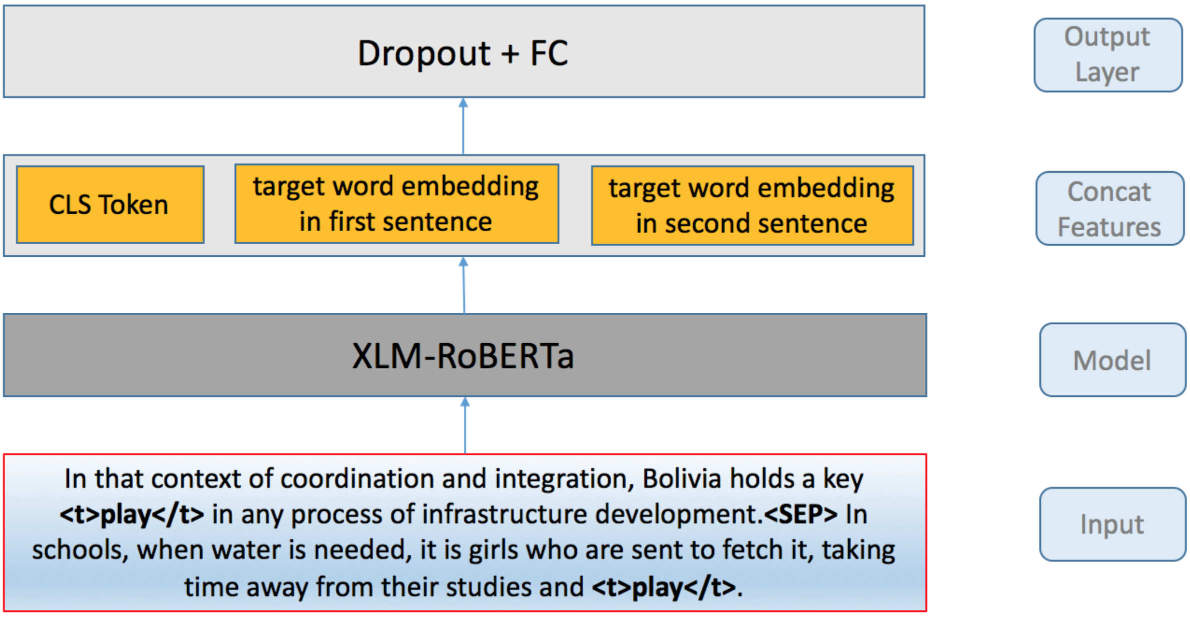
Dropout + FC
Output Layer
CLS Token
target word embedding in first sentence
target word embedding in second sentence
Concat Features
XLM-RoBERTa
Model
In that context of coordination and integration, Bolivia holds a key <t>play</t> in any process of infrastructure development.<SEP> In schools, when water is needed, it is girls who are sent to fetch it, taking time away from their studies and <t>play</t>.
Input

Figure 2: Fine-tuned XLM-RoBERTa model architecture.

	Training	Test
No. of target words	3,726	491
No. of pairs	8,000	1,000
Min. tokens	6	5
Avg. tokens (original)	24	26
Max. tokens (original)	88	116
Max. tokens (post-proc.)	81	81

Table 1: Statistics of the data

unique id of the pair, target lemma, part-of-speech in {NOUN, VERB, ADJ, ADV}, the first sentence, the second sentence, the start and the end indices (zero-based numbering) of the target word in the first and the second context, respectively. The .gold file contains unique id of the pair and the label, which is represented by T or F.

For the training dataset, we clean up the text by completing word abbreviation, removing special punctuation, and segmenting the sentences into subword lists by Byte-Pair Encoding (BPE) (Sennrich et al., 2015). Since it is difficult to capture the meaning of the target word in the context for long sentences (Pan et al., 2019; Zhu et al., 2021), we limit the length of each sentence with maximum 40 words before and after the target word.

We include additional resource, WordNet, to augment our training data because WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concepts. Here, we randomly select example sentences of target word in WordNet to expand our training corpus, which increases around 30% of training data. By such preprocessing, we obtain the dataset and report the statistics in Table 1.

3.3 Model Design

Figure 2 outlines our model architecture, which consists of four modules, i.e., input design, model learning, final feature construction, and the classifier. The whole framework is based on fine-tuning the pre-trained XLM-RoBERTa model to conduct binary classification on two given contexts. Different from the inputs for XLM-RoBERTa, the input of our system contains of the following modifications: first, in order to highlight the target word in the contexts, we borrow the setting in (Lei et al., 2017; Botha et al., 2020) by adding special symbols <t> and </t> to embrace the target word in the contexts. Given the example presented in Fig. 1, the target word of "play" is then embraced by the additional symbols, <t> and </t> in the contexts. Second, we concatenate the given two contexts by <SEP>. Figure 2 illustrates the result in the input module. Moreover, in the experiment, we exchange the order of the contexts to get more training data.

After learning the tokens' representations by XLM-RoBERTa, we construct a new vector by concatenating the [CLS] token's representation in the last layer of XLM-RoBERTa and the representations of the target word in both sentences. As BPE tokenization may separate a target word into several subwords, we compute its representation by averaging the corresponding representations. Next, the newly constructed feature is fed into a fully-connected network to compute the final binary prediction probability.

During training, we conduct the following techniques to increase the model convergence and robustness:

- **Optimizer.** We adopt the Ranger (Yong et al., 2020) optimizer to replace the AdamW because it is a more synergistic optimizer combining rectified Adam and the look-ahead mechanism with gradient centralization in one optimizer.
- **Adversarial training.** We apply the fast gradient method (Miyato et al., 2017) in the training to obtain more stable word representations.
- **Cross validation.** We also apply stratified K-fold cross validation on the training set and the development set. For each fold, we hold the group as a local test set and set the remaining groups as the training set. We then average the model prediction on each fold as the final prediction to obtain more robust results.
- **Pseudo labelling.** Pseudo labelling (Wu and Prasad, 2018) is an effective semi-supervised learning method to utilize the abundant unlabeled data via their pseudo labels. In this work, we first train our model on the training set. Next, we apply the trained model to predict the En-En multi-lingual test set and use the predicted labels as our pseudo labels. Finally, both the training set and the En-En pseudo labels are included to train a final model. Especially, we observe that by this trick, this final model can improve the prediction performance on cross-lingual tasks slightly.

It is noted that in the cross-lingual tasks, we do not back-translate the subwords to English but apply the same model trained from the En-En dataset because it allows us to maintain the target word in the corresponding languages seamlessly. This is similar to the procedure in (Artetxe et al., 2020).

4 Experiments

In the following, we detail our experimental setup and present the results with analysis.

4.1 Setup

Our code is written in Pytorch based on the Huggingface Transformer library [1] for XLM-RoBERTa. Other hyperparameters are set based on our hand-on experience. For example, the seed for the random generator is set to 3,999. The batch size is set to 10 and the hidden feature size is 1,024. The maximum length limit of a context is 240 though it is unreachable because we have conducted trimming in the data preprocessing procedure. The dropout rate is tested from $\{0.2, 0.3, 0.25, 0.28\}$ and finally fixed to 0.28. K in the stratified cross validation is set to 5. The two special tokens, <t> and </t>, are included into the word dictionary for learning.

The training data consists of the official En-En multilingual training corpus and the contexts from WordNet. At the beginning, we choose XLM-RoBERTa$_{\text{Base}}$ as the backbone of our system to explore the possibility of our implementation tricks. After identifying the effectiveness of the designed input in Sec. 3.2, we apply XLM-RoBERTa$_{\text{Large}}$ to tune the corresponding hyperparameters, such as changing the learning rate, the batch size, the dropout rate, and the early stop mechanism. Furthermore, we observe that long contexts may ignore the importance of the target word in the contexts. Hence, we center on the target words to cut off the contexts at both ends with a certain length. To further strengthen the influence of the target word in a context, we concatenate the embedding of the [CLS] token with the embeddings of the target word in the contexts as the final input for the logit fully-connected network. From our experiment, this strategy can significantly boost the model performance while improving the convergence.

4.2 Results

Table 2 reports the results of different implementation strategies on the tasks. From the results, we observe that

- By replacing XLM-RoBERTa$_{\text{Base}}$ with XLM-RoBERTa$_{\text{Large}}$, we can gain at least 3% improvement on all tasks.
- By applying Ranger optimizer, we attain the results in Large+RO, which gain an average increase of 0.2% per task. We conjecture the improvement comes from the fact that the model converges to a more optimal solution.
- In Large+RO+LRA, we vary the learning rate

[1] https://github.com/huggingface/transformers

Strategy	Avg	En-En	Fr-Fr	Ru-Ru	Zh-Zh	Ar-Ar	En-Ru	En-Zh	En-Fr	En-Ar
Base	80.8	85.5	80.7	78.7	80.9	79.1	81.0	81.9	79.1	80.4
Large	85.1	88.2	84.2	84.3	87.0	82.6	84.6	85.7	85.6	83.9
Large + RO	85.4	88.7	85.3	85.1	86.9	83.3	84.7	85.9	84.7	83.6
Large + RO + LRA	85.4	89.2	84.9	85.1	86.8	83.2	84.8	85.6	85.2	84.1
Large + RO + LRA + ES	85.5	89.0	84.8	85.5	86.9	83.1	85.2	85.2	85.4	84.1
Large + RO + CTWE	86.3	90.0	85.8	85.9	87.1	83.9	86.0	86.1	85.4	86.2
Large + RO + CTWE + HC	86.3	89.9	85.7	85.9	87.2	84.2	85.6	86.7	85.0	86.3
Large + RO + CTWE + HC + WordNet	86.5	91.6	85.8	85.7	87.1	84.0	85.3	87.1	85.6	86.0
Large + RO + CTWE + HC + WordNet + AT	87.0	91.1	86.3	85.9	87.9	85.1	86.3	87.2	86.3	86.9
Large + RO + CTWE + HC + WordNet + AT + PL	**88.1**	**91.7**	**86.9**	**86.5**	**89.2**	**86.5**	**88.0**	**87.9**	**88.6**	**87.2**

Table 2: Results of fine-tuning XLM-RoBERTa under different strategies. The abbreviation is defined as follows: Base: XML-RoBERTa$_{\text{Base}}$; Large: XML-RoBERTa$_{\text{Large}}$; RO: Ranger Optimizer; LRA: learning rate adjustment; ES: early stop; CTWE: concatenating target words' embeddings; HC: the best parameters for LRA and ES; AT: adversarial training; PL: pseudo labels.

from 1.5e-5, 1.3e-5, 1.2e-5, to 1.21e-5 progressively and finally find that when the learning rate is 1.2e-5, we attain the best performance. We then search the optimal epoch for the early stop by setting the maximum number of epoch to 10. In Large+RO+LRA+ES, we observe the optimal epoch for early stop (the patience value) is 3. These parameters are then fixed for HC. From the results, we notice that tuning the learning rate and adopting the early stop mechanism can improve the model performance accordingly.

- By concatenating target the word embedding, we obtain the results in Large+RO+CTWE, and actually, our model can be trained with fewer epochs and attain around 1.1% improvement on average.
- By adding more training data from WordNet, we get another 0.2% average improvement in Large+RO+CTWE+HC+WordNet. We conjecture the improvement mainly comes from the increase of the training data.
- By adding the Pseudo label data, we can gain another 0.8% average improvement. The score of EN-EN test dataset is generally higher than other test dataset. We discover that the first 462 pieces of English test dataset have the same target word as test dataset in other tasks. Therefore, adding EN-EN pseudo label helps predict other tasks.

In sum, we conclude that by applying XLM-RoBERTa$_{\text{Large}}$ on the Ranger optimizer, the target word embedding concatenation mechanism, more external training data, and pseudo labels, we can improve the model performance accordingly.

Finally, our system attains the champion on the En-Ar, En-Fr, En-Ru, and En-Zh cross-lingual tasks. In multilingual tasks, we also sit at eighth place, seventh place, sixth place, seventh place, and fifth place for the En-En, Ar-Ar, Fr-Fr, Ru-Ru, Zh-Zh tasks, respectively.

5 Conclusion

In this paper, we present our system to tackle the word-in-context disambiguation task. We fine-tune the XLM-RoBERTa model to solve both multilingual and cross-lingual word-in-context disambiguation tasks. We specifically design the input format to emphasize the target word in two contexts and promote the importance of the target word by concatenating the embeddings in the corresponding context with the [CLS] token to output the classification probability. We apply several training tricks to improve the robustness of model and attain improvement during this procedure. The competition results demonstrate the effectiveness of our implementation. In the future, we plan to explore more model architecture to boost the performance for multilingual tasks.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *ACL 2020*, pages 4623–4637. Association for Computational Linguistics.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity linking in 100 languages. In *EMNLP*, pages 7833–7845. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *CoRR*, abs/1705.00440.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of tricks for efficient text classification. In *EACL*, pages 427–431. Association for Computational Linguistics.

Wenqiang Lei, Yisong Miao, Runpeng Xie, Bonnie Webber, Meichun Liu, Tat-Seng Chua, and Nancy F Chen. 2021. Have we solved the hard problem? it's not easy! contextual lexical contrast as a means to probe neural coherence. In *AAAI*.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: a simple word interaction model for implicit discourse relation recognition. In *IJCAI*, pages 4026–4032.

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. *CoRR*, abs/1907.11692.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *ICLR*. OpenReview.net.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Trans. Assoc. Comput. Linguistics*, 2:231–244.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Mohammad Taher Pilehvar and José Camacho-Collados. 2018. Wic: 10,000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *ICLR*. OpenReview.net.

Xinyi Wang, Haiqin Yang, Liang Zhao, Yang Mo, and Jianping Shen. 2021. Refbert: Compressing bert by referencing to pre-computed representations. In *IJCNN*.

Hao Wu and Saurabh Prasad. 2018. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Trans. Image Process.*, 27(3):1259–1270.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *NeurIPS*.

Haiqin Yang and Jianping Shen. 2021. Emotion dynamics modeling via bert. In *IJCNN*.

Haiqin Yang, Xiaoyuan Yao, Yiqun Duan, Jianping Shen, Jie Zhong, and Kun Zhang. 2021. Progressive open-domain response generation with multiple controllable attributes. In *IJCAI*.

Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. 2020. Gradient centralization: A new optimization technique for deep neural networks. In *ECCV*, volume 12346 of *Lecture Notes in Computer Science*, pages 635–652. Springer.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.