IBERT: Idiom Cloze-style reading comprehension with Attention

Anonymous EMNLP submission

Abstract

Idioms are permanent phrases that are often formed from tales. They are prevalent in informal discussions and literary works. Their meanings are often quite devoid of composition. The idiom cloze task is a difficult research challenge in Natural Language Processing (NLP). On available datasets, sequence-to-sequence (Seq2Seq) model-based approaches to this problem fared pretty well. However, they lack comprehension of the non-compositional nature of idiomatic idioms. In addition, they do not evaluate both the local and global contexts simultaneously. In this research, we present a BERT-based embedding Seq2Seq model that captures idiomatic phrases and takes global and local contexts into account. Our methodology uses XLNET as the encoder and Roberta to choose the most likely idiom for a given scenario. Experiments conducted on the EPIE Static Corpus dataset demonstrate that our approach outperforms the current state-of-the-art.

1 Introduction

The cloze test is a test in which participants are asked to complete a paragraph using the appropriate words. Typically, cloze exams assess a participant's ability to grasp a particular material. The cloze task differs significantly from other Natural Language Processing (NLP) tasks in that it demands a much greater long-term memory to make judgments and interpret the information. The completion of these comparative examinations will shed light on present NLP tasks in text comprehension. This study solves the cloze problem using a BERT-based sequence-to-sequence (Seq2Seq) model. Given a particular context in the form of a passage, two stages are required to develop the answer to a cloze issue based on this context. The first is to comprehend the meaning of the phrase, and the second is to choose the appropriate idiom for each "blank" - where the original word has been omitted – in the sentence. Utilizing neural network models such as Knowledgeable Reader (Mihaylov and Frank, 2018) and Entity Tracking (Hoang et al., 2018), prior research has approached NLP challenges comparable to the cloze task using neural network techniques. These techniques handle idioms as though they were ordinary terms. Nevertheless, idiomatic idioms often contain meanings that are very noncompositional and should not be taken literally. The saying "it's raining cats and dogs," for instance, implies cats and dogs dropping from the sky if interpreted literally. This notion is inaccurate, since this term is often used to indicate severe rain. Failure to comprehend the proper meaning of colloquial terms is damaging to the models' decision-making process. Even if a model is capable of comprehending the paragraph, it cannot connect a phrase that is completely unrelated to the context from a literal standpoint. This necessitates other strategies for the cloze problem as our paradigm. It necessitates offering a grammatical candidate and maintaining semantic coherence. Earlier work on contextual connection comprehension has produced solutions with outstanding performance. For example, the BERT (Devlin et al., 2019) is effective in comprehending the context of a specific situation. In addition, pre-trained BERT and related models may be customized to perform contextual comprehension for other activities. In this research, we solve the context comprehension stage of the cloze problem using pre-trained models. To accurately comprehend the meaning of idiomatic idioms, we deploy another pre-trained XLNET (Yang et al., 2019) model and fine-tune it to tackle the cloze problem. We teach XLNET the local context of the deleted word (the phrase from which the original word was removed) and the global context of the removed word (the entire passage). This information is provided by the last layer output of XLNET's hidden layers. Combining the contextual embeddings provided by our BERT-based pretrained models with the idiom embeddings from the XLNET model. Softmax (Liu et al., 2016) is applied on the aggregated embeddings to choose the optimal word. Our models are calibrated using the EPIE (Saxena and Paul, 2020) Static Corpus data collection.

2 Related Work

2.1 Cloze-style reading comprehension

Cloze-style reading comprehension utilizes a passage of word tokens $x_{1:n}$ with one token x_j masked; the aim is to replace the masked word y, which was originally at position j, with the correct word. Numerous works have already attained outstanding results in the cloze test (Mihaylov and Frank, 2018; Hoang et al., 2018; Schick and Schütze, 2021). Researchers have produced

many large-scale cloze-style reading comprehension datasets, including RACE (Lai et al., 2017) and Children's Book Test (CBT) (Hill et al., 2015). However, these studies only test regular words for Cloze-style reading comprehension, and idiom phrases are often out of context inside the paragraphs. In this research project, we want to use cutting-edge technology in order to achieve cutting-edge performance in Idiom Cloze-style reading comprehension. The EPIE dataset utilized in this article is likewise a large-scale cloze-style dataset, but it focuses on idiom prediction in English.

2.2 Pre-trained Language Models

Massive sources of unilateral context may affect the model's accuracy throughout regular use. In reality, it is quite probable that they will lack symbols following sentences and polysemous conditions inside the phrase. There are earlier studies on enhancing word embedding (Mikolov et al., 2013; Peters et al., 2018), but they do not assist us in resolving the difficulty of our jobs. With the introduction of transformer (Vaswani et al., 2017), numerous pre-trained models like as BERT and XLNet were offered in the NLP field. Language model pre-training has been shown to be successful on a list of natural language tasks at both the sentence-level (Bowman et al., 2015) and token-level (Tjong Kim Sang and De Meulder, 2003), according to several studies.

2.3 Idiom detection

The reading comprehension of idioms is the subject of a substantial amount of current study. Dual Embedding Model with bert (Tan and Jiang, 2020) was the first popular neural network reading comprehension model. In this prior study by Tan et al. (Tan and Jiang, 2020), the contextual sentence and candidate phrases are encoded using the BERT model. However, the basic BERT model lacks confidence in common sense and pragmatic inference and is incapable of processing lengthy text sequences. In addition, the BERT model performs poorly in the negation scenario of the statement. For others, Cloze-style reading comprehension comprises of a substantial context size, and the masks' matching keys are far away. In addition, we were unable to disregard the negation and common sense conditions in the Cloze-style reading comprehension test.

3 Method

3.1 Task Definition

Our project's fundamental concept is the idiom cloze test. We choose an idiom as a contender and then eliminate it from each phrase. For these candidates, each blank space must be filled up. Therefore, the first step is to pad each candidate. Then we must locate the best qualified individual whenever a vacancy arises.

3.2 Padding Sequence

This section begins with candidate processing. The initial phase of our endeavor is to comprehend idioms or candidates. In the subsequent phases, we will look for the best qualified applicant. In order to search, we must process these candidates.

This is necessary since each candidate has a unique length, making it difficult to perform any operation on them. The default padding and most common padding is zero-padding, in which we add zeros to the rear of each candidate. We will utilize the length of the candidate with the longest length as the padding length. This may be accomplished by invoking specified functions. We must initially look for the candidate with the greatest length.

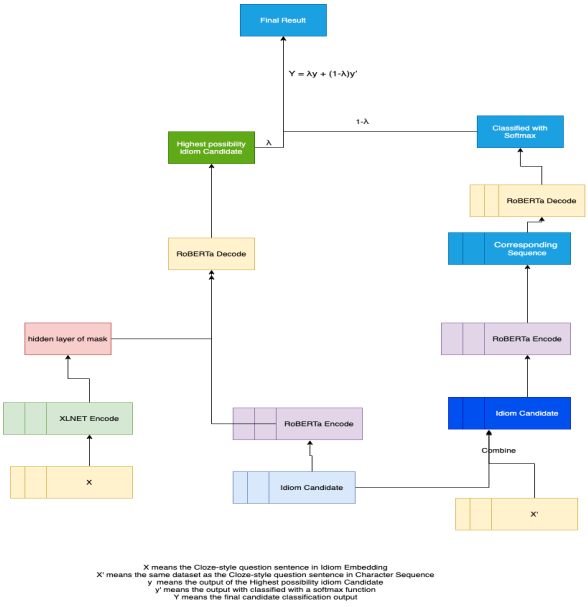

Figure 1: Our Method Diagram about how our Attention Baselines works

3.3 Attention Baselines

Previous approaches used to BERT-based Dual Embedding were based only on the BERT architecture. Due to the weakness in BERT and the success of XLNET for various NLP tasks, including reading comprehension, we suggested to offer new approaches (shown in Figure 1) based on XLNET and BERT to address Idiom Cloze-style reading comprehension. In the first, an English idiom is treated as a series of letters, and the BERT is used to examine the original link between each contextual word. Then, we aggregate the sections containing each potential phrase into numerous sequences, one for each contender. In the second baseline, each idiom is treated as a single token accompanied by its embedding vector. We utilize XLNET to analyze the text and then compare the encoded passage with the embedding of each possible idiom.

Attention Baseline with Idioms as Character Sequence Attention baseline with idiom as candidate sequence is work required to implement the BERT model for idiom cloze-style comprehension.Given a passage P = (p_1, p_2, p_3, $\cdots$, $[MASK]$, $\cdots$, p_n) and a

candidate $d_k \in D$, we concatenate them into a single sequence ([CLS], p_1, p_2, p_3, $\cdots$,d_{n_1},d_{n_2}, $\cdots$,d_{n_k}, $\cdots$,p_n,[SEQ]), where d_{n_1} to d_{n_k} are the characters and padding of the idiom d_n. We may use the BERT to immediately analyze this sequence and extract the hidden representation for [CLS] in the final hidden layer, denoted by $h^L_{k,0} \in R^d$. To find the candidate idiom d_k among all candidates, we use the linear layer to process $h^L_{k,0}$ for $k = 1, 2, \cdots, K$ and the softmax function with each candidate's probability value in D. Then, we will choose the best option for our cloze-style reading comprehension as the final selection.

Attention Baseline with Idiom Embedding Numerous idioms are non-compositional; their meaning cannot be inferred simply from their component characters. For instance, "It's a piece of cake" actually implies multiple cakes, but it is often used to refer to a work that is relatively simple. Therefore, if we merely embed the meaning of each character, this might result in considerable confusion. A single embedding vector for the full idiom, on the other hand, may assist the model comprehend the contextual connection in reading comprehension.

In this baseline, instead of concatenating the passage and a potential response, we divide them into a single BERT model sequence. We processed the passage sequence using XLNET to ([CLS], p_1,p_2,p_3, $\cdots$, $[MASK]$, $\cdots$,p_n,[SEP]). Then, we utilize the h^L_b hidden representation of [MASK] at the decoding layer to match each candidate's response with BERT. This manner, regardless of the number of contenders, we only embed the whole section once using the XLNET paradigm. It may assist our work avoid the issue of non-compositional idioms.

The embedding vector each candidate $d_k \in D$ is denoted by d_k, and the hidden representation h^L_b is compatible with each candidate idiom by element-wise multiplication. The probability of choosing d_k from among all the candidates is thus determined by the equation 1.

$$p_k = \frac{exp(w \cdot (d_k \otimes h^L_b) + b)}{\sum_{d'=1}^{D} exp(w \cdot (d_k \otimes h^L_b) + b)} \quad (1)$$

$w \in R^d$ and $b \in R$ are parameters of the model, and $\otimes$ is element-wise multiplication. We utilize cross-entropy loss as the loss function to train the model.

3.4 Context-aware Pooling

The baseline for the attention model contains possible flaws. It is straightforward to note that idioms are never composed. In order for the proposed idiom to fit well in the passage, not only must its grammar correspond to the surrounding context, but its meaning must also match the paragraph incredibly well. To solve the cloze-style reading comprehension, however, it is necessary that we understand how to induce the applicant to use more complex language, such as verbs and nouns. In order for our baseline to be context-aware, we need additional work.

As the milestone of the transformer, increasingly more pre-train models enable the context-aware function with the contextual environment. In addition, several models enable global context-awareness, such as the BERT work in the SQuAD dataset (Rajpurkar et al., 2016). Therefore, we determine if an idiom candidate is appropriate for a paragraph. In addition to understanding its neighboring contextual facts, we must also comprehend the passage's semantic significance. In addition to matching the idiom candidate to its context, we must also match the idiom's meaning to the whole paragraph. Recall that $H^L = (h^L_0, h^L_1, h^L_2, \cdots, h^L_n)$ represents the hidden states of the last hidden layer of baseline after the sequence is processed. Our model with context-aware pooling may function similarly to the equation 2.

$$p_k = \frac{exp(d_k \cdot h^L_b + max^n_{i=0}(d_k \cdot h^L_i))}{\sum_{d'=1}^{D} exp(d_k \cdot h^L_b + max^n_{i=0}(d_k \cdot h^L_i))} \quad (2)$$

3.5 Dual Pretrain Attention Model

In our technique (Figure 1), we suggested using linear interpolations (Zhang et al., 2018) to enhance our idiom identification solution for cloze-style reading comprehension smoother. We execute linear interpolations in the final texture hidden space between both training outputs, one from the output categorized with the softmax function in Attention Baseline with Idioms as Character Sequence and the other from Attention Baseline with Idiom Embedding. We employ the $lambda$ parameter as a weight to guarantee that both baselines can operate efficiently on the final result. In our model, $lambda$ represents a beta distribution sample. Therefore, we guarantee that lambda is greater than 0.7 and that the combination is dominated by the Attention Baseline with Idioms as Character Sequence.

With linear interpolation, we may rapidly get a new candidate option, and for cloze-style reading comprehension, we choose the candidate with the greatest score.

4 Data

We test the accuracy of the classification model using a single standard dataset - EPIE, which contains 359 candidate types (given in table 1). Then, we develop a classification software to determine which idioms should be utilized in cloze-style reading comprehension phrases. In our program, we designed a self-attention program for the job. Then, we use the training, validation, and test divides described by Lee and Dernoncourt (Lee and Dernoncourt, 2016) to examine our classification loss score.

As seen in Table 1, we have a total of 21890 candidates. Each candidate has an associated tag (displayed as 2). Both O, B-IDIOM, and I-IDIOM denote a single word or special symbol position. Given a single statement, we must eliminate the portion beginning with "B-IDIOM" and ending with "I-IDIOM." The 21890 sentences were obtained by removing the candidate from the original sentences.

We use three kinds of data labels. To label our data, we prepend ['CLS'] to the beginning and ['SEP'] to the end of each of the 21890 candidates. To name our deleting sentences, we append ['CLS'] to the beginning and ['SEP'] to the conclusion of each phrase in which the candidate is deleted. In addition, a ['UNK'] label is added to the location of the candidate for deletion in the 21890 sentences. By labeling the preceding, it would be possible to encode all 21891 possibilities and eliminate sentences. The sample is shown in the table 3.

In which, |T| represents the number of idiom candidate and |N| represents the sentence data size.

Table 1: Number of Sentences in the Dataset

Dataset	Train	Validation	Test	T	N
EPIE	15k	5k	2k	359	22k

Table 2: Type of Tags in the Dataset

Type1	Type2	Type3
O	B-IDIOM	I-IDIOM

Table 3: Example of Encoding and Labeling

Original Sentence:Anyway , thanks MKM and keep up the good work !
Candidate:keep up the good work
Label Candidate:'[CLS]', 'keep', 'up', 'the', 'good', 'work', '[SEP]'
encode:11815, 17, 19, 3466, 414, 536, 692, 21, 435, 76, 18, 195, 154, 17, 136, 4, 3
deleting sentence:'Anyway', ',', 'thanks', 'MKM', 'and', '!

5 Result

Table 4 compares the classification accuracy of our technique to that of various other models. As a baseline, we use the Dual BERT Embedding Model with Entity Tracking. Some ways to classifying the idiom candidate for cloze-style reading comprehension make use of self-attention for methods using attention and in-depth contextualization of word representation. However, their performance was inferior to that of our model. Each model and its variables were trained eight times, resulting in an average performance.

As seen by our experiment in Figure reffig:epoch, our morel loss decreases as the number of epochs increases. In addition, we collaborate with the other two types of baselines and our approach utilizing the precision (shown in the table 4). As can be shown, our technique has 7.21 % greater accuracy than the Dual BERT Embedding Model and 14.66 % better accuracy than Entity Tracking. It makes sense that the Dual BERT

Figure 2: Epoch number and the loss

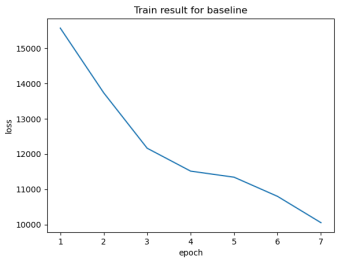

Table 4: Accuracy of task in Idiom Cloze-style Reading comprehension Performance with baselines

Model	EPIE(%)
Tan et al. (Tan and Jiang, 2020)	71.02
Our Method	78.23
Entity Tracking (Hoang et al., 2018)	63.57

Embedding Model is intended to address the difficulty of Chinese Idiom Cloze-style reading comprehension, while Entity Tracking is just intended for the standard Cloze-style Reading comprehension. The Idiom is often non-compositional and is rather sophisticated. From the study of Tan et al., we can conclude that their Dual BERT Embedding Model has a superior performance on Chinese Idiom Cloze-style Reading comprehension. They already achieve 84.43 % under Chinese settings.

6 Conclusion and Future work

We created a novel model that meticulously executed the Idiom Cloze-style reading comprehension task and compared it to commonly-used algorithms using the EPIE dataset. Using several word representation approaches, we discovered that context information continue to have a significant impact on classification performance. Our technique performs 7.21% better on the EPIE dataset challenge of cloze-style reading comprehension than the Dual BERT Embedding Model. Due of time constraints, we only focus on one dataset for our understanding. However, overcoming the obstacle of idiom cloze-style reading comprehension is a new milestone for us.

In future research, we will investigate other attention mechanisms, including block self-attention (Shen et al., 2018), hierarchical attention (Yang et al., 2016), and hypergraph attention (Bai et al., 2021). These methods may combine information from several positional representations and capture both local and long-range context dependencies. In addition, we intend to do more experiments on several idiom datasets, such as CHID (Zheng et al., 2019) and LIDIOMS (Moussallem et al., 2018). With these additional types of datasets, we want to do more testing and develop more robust algorithms for solving the Idiom Cloze-style reading comprehension challenge.

References

Song Bai, Feihu Zhang, and Philip H.S. Torr. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. 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 *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, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Luong Hoang, Sam Wiseman, and Alexander Rush. 2018. Entity tracking improves cloze-style reading comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1049–1055, Brussels, Belgium. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 507–516. JMLR.org.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIDIOMS: A multilingual linked idioms data set. *CoRR*, abs/1802.08148.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue*, pages 87–94, Cham. Springer International Publishing.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. 2018. Bi-directional block self-attention for fast and memory-efficient sequence modeling. *arXiv preprint arXiv:1804.00857*.

Minghuan Tan and Jing Jiang. 2020. A BERT-based dual embedding model for Chinese idiom prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1312–1322, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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*, volume 32, pages 5753–5763. Curran Associates, Inc.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.