ERNIE: Enhanced Representation through Knowledge Integration

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng
Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu
Baidu Inc.
{sunyu02,wangshuohuan,liyukun01,fengshikun01,tianhao,wu_hua}@baidu.com

Abstract

We present a novel language representation model enhanced by knowledge called ERNIE (Enhanced Representation through kNowledge IntEgration). Inspired by the masking strategy of BERT (Devlin et al., 2018), ERNIE is designed to learn language representation enhanced by knowledge masking strategies, which includes entity-level masking and phrase-level masking. Entity-level strategy masks entities which are usually composed of multiple words. Phrase-level strategy masks the whole phrase which is composed of several words standing together as a conceptual unit. Experimental results show that ERNIE outperforms other baseline methods, achieving new state-of-the-art results on five Chinese natural language processing tasks including natural language inference, semantic similarity, named entity recognition, sentiment analysis and question answering. We also demonstrate that ERNIE has more powerful knowledge inference capacity on a cloze test.

1 Introduction

Language representation pre-training (Mikolov et al., 2013; Devlin et al., 2018) has been shown effective for improving many natural language processing tasks such as named entity recognition, sentiment analysis, and question answering. In order to get reliable word representation, neural language models are designed to learn word co-occurrence and then obtain word embedding with unsupervised learning. The methods in Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) represent words as vectors, where similar words have similar word representations. These word representations provide an initialization for the word vectors in other deep learning models. Recently, lots of works such as Cove (McCann et al., 2017), Elmo (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) improved word representation via different strategies, which has been shown to be more effective for down-stream natural language processing tasks.

The vast majority of these studies model the representations by predicting the missing word only through the contexts. These works do not consider the prior knowledge in the sentence. For example, In the sentence " *Harry Potter is a series of fantasy novels written by J. K. Rowling*". *Harry Potter* is a novel name and *J. K. Rowling* is the writer. It is easy for the model to predict the missing word of the entity *Harry Potter* by word collocations inside this entity without the help of long contexts. The model cannot predict *Harry Potter* according to the relationship between *Harry Potter* and *J. K. Rowling*. It is intuitive that if the model learns more about prior knowledge, the model can obtain more reliable language representation.

In this paper, we propose a model called ERNIE (enhanced representation through knowledge integration) by using knowledge masking strategies. In addition to basic masking strategy, we use two kinds of knowledge strategies: phrase-level strategy and entity-level strategy. We take a phrase or a entity as one unit, which is usually composed of several words. All of the words in the same unit are masked during word representation training, instead of only one word or character being masked. In this way, the prior knowledge of phrases and entities are implicitly learned during the training procedure. Instead of adding the knowledge embedding directly, ERNIE implicitly learned the information about knowledge and longer semantic dependency, such as the relationship between entities, the property of a entity and the type of a event, to guide word embedding learning. This can make the model have better generalization and adaptability.

In order to reduce the training cost of the model, ERNIE is pre-trained on heterogeneous Chinese data, and then applied to 5 Chinese NLP tasks. ERNIE advances the state-of-the-art results on all of these tasks. An additional experiment on the cloze test shows that ERNIE has better knowledge inference capacity over other strong baseline methods.

Our Contribution are as follows:

(1) We introduce a new learning processing of language model which masking the units such as phrases and entities in order to implicitly learn both syntactic and semantic information from these units.

(2) ERNIE significantly outperforms the previous state-of-the art methods on various Chinese natural language processing tasks.

(3) We released the codes of ERNIE and pre-trained models, which are available in https://github.com/PaddlePaddle/LARK/tree/develop/ERNIE.

2 Related Work

2.1 Context-independent Representation

Representation of words as continuous vectors has a long history. A very popular model architecture for estimating neural network language model (NNLM) was proposed in (Bengio et al., 2003), where a feed forward neural network with a linear projection layer and a non-linear hidden layer was used to learn the word vector representation.

It is effective to learn general language representation by using a large number of unlabeled data to pretrain a language model. Traditional methods focused on context-independent word embedding. Methods such as Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) take a large corpus of text as inputs and produces a word vectors, typically in several hundred dimensions. They generate a single word embedding representation for each word in the vocabulary.

2.2 Context-aware Representation

However, a word can have completely different senses or meanings in the contexts. Skip-thought (Kiros et al., 2015) proposed a approach for unsupervised learning of a generic, distributed sentence encoder. Cove (McCann et al., 2017) show that adding these context vectors improves performance over using only unsupervised word and character vectors on a wide variety of common NLP tasks. ULMFit (Howard and Ruder, 2018) proposed an effective transfer learning method that can be applied to any task in NLP. ELMo (Peters et al., 2018) generalizes traditional word embedding research along a different dimension. They propose to extract context-sensitive features from a language model. The GPT (Radford et al., 2018) enhanced the context-sensitive embedding by adapting the Transformer.

BERT (Devlin et al., 2018) uses two different pretraining tasks for language modeling. BERT randomly masks a certain percentage of words in the sentences and learn to predict those masked words. Moreover, BERT learn to predict whether two sentences are adjacent. This task tries to model the relationship between two sentences which is not captured by traditional language models. Consequently, this particular pretraining scheme helps BERT to outperform state-of-the-art techniques by a large margin on various key NLP datasets such as GLUE (Wang et al., 2018) and SQUAD (Rajpurkar et al., 2016) and so on.

Some other researchers try to add more information based on these models. MT-DNN (Liu et al., 2019) combine pre-training learning and multi-task learning to improve the performances over several different tasks in GLUE (Wang et al., 2018). GPT-2 (Radford et al., 2019) adds task information into the pre-training process and adapt their model to zero-shot tasks. XLM (Lample and Conneau, 2019) adds language embedding to the pre-training process which achieved better results in cross-lingual tasks.

2.3 Heterogeneous Data

Semantic encoder pre-trained on heterogeneous unsupervised data can improve the transfer learning performance. Universal sentence encoder (Cer et al., 2018) adopts heterogeneous training data drawn from Wikipedia, web news, web QA pages and discussion forum. Sentence encoder (Yang et al., 2018) based on response prediction benefits from query-response pair data drawn from Reddit conversation. XLM (Lample and Conneau, 2019) introduce parallel corpus to BERT, which is trained jointly with masked language model task. With transformer model pre-trained on heterogeneous data, XLM shows great performance gain on supervise/unsupervised MT task and classification task.

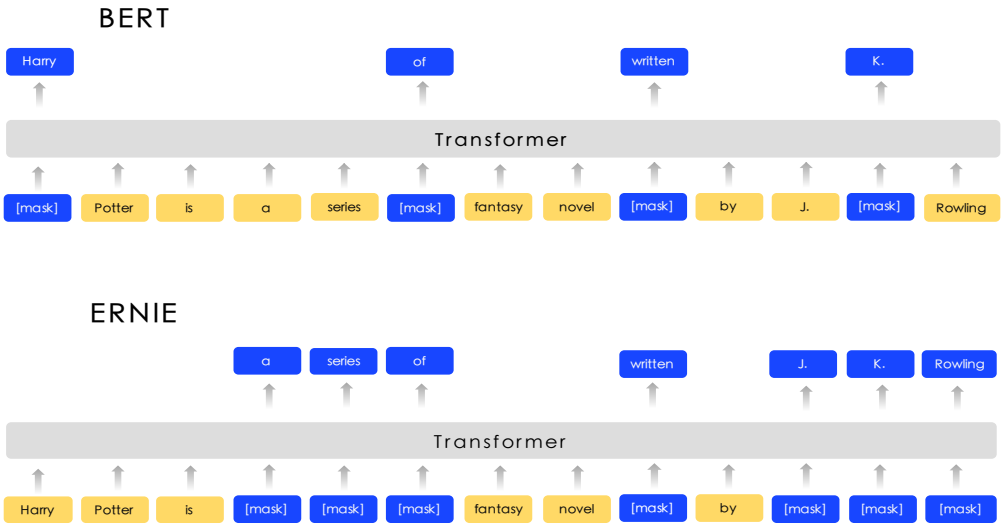

Figure 1: The different masking strategy between BERT and ERNIE

3 Methods

We introduce ERNIE and its detailed implementation in this section. We first describe the model's transformer encoder,and then introduce the knowledge integration method in Section 3.2. The comparisons between BERT and ERNIE are shown visually in Figure 1.

3.1 Transformer Encoder

ERNIE use multi-layer Transformer (Vaswani et al., 2017) as basic encoder like previous pretraning model such as GPT, BERT and XLM. The Transformer can capture the contextual information for each token in the sentence via self-attention, and generates a sequence of contextual embeddings.

For Chinese corpus, we add spaces around every character in the CJK Unicode range and use the WordPiece (Wu et al., 2016) to tokenize Chinese sentences. For a given token, its input representation is constructed by summing the corresponding token, segment and position embeddings. The first token of every sequence is the special classification embedding([CLS]).

3.2 Knowledge Integration

we use prior knowledge to enhance our pretrained language model. Instead of adding the knowledge embedding directly, we proposed a multi-stage knowledge masking strategy to integrate phrase and entity level knowledge into the Language representation. The different masking level of a sentence is described in Figure 2.

3.2.1 Basic-Level Masking

The first learning stage is to use basic level masking, It treat a sentence as a sequence of basic Language unit, for English, the basic language unit is word, and for Chinese, the basic language unit is Chinese Character. In the training process, We randomly mask 15 percents of basic language units, and using other basic units in the sentence as inputs, and train a transformer to predict the mask units. Based on basic level mask, we can obtain a basic word representation. Because it is trained on a random mask of basic semantic units, high level semantic knowledge is hard to be fully modeled.

3.2.2 Phrase-Level Masking

The second stage is to employ phrase-level masking. Phrase is a small group of words or characters together acting as a conceptual unit. For English, we use lexical analysis and chunking tools to get the boundary of phrases in the sentences, and use some language dependent segmentation tools to get the word/phrase information in other language such as Chinese. In phrase-level mask stage, we also use basic language units as training input, unlike random basic units mask, this time we randomly select a few phrases in the sentence, mask and predict all the basic units in the same phrase. At this stage, phrase information is encoded into the word embedding.

3.2.3 Entity-Level Masking

The third stage is entity-level masking. Name entities contain persons, locations, organizations, products, etc., which can be denoted with a proper

Figure 2: Different masking level of a sentence

name. It can be abstract or have a physical existence. Usually entities contain important information in the sentences. As in the phrase masking stage, we first analyze the named entities in a sentence, and then mask and predict all slots in the entities. After three stage learninga word representation enhanced by richer semantic information is obtained.

4 Experiments

ERNIE was chosen to have the same model size as BERT-base for comparison purposes. ERNIE uses 12 encoder layers, 768 hidden units and 12 attention heads.

4.1 Heterogeneous Corpus Pre-training

ERNIE adopts Heterogeneous corpus for pre-training. Following (Cer et al., 2018), we draw the mixed corpus Chinese Wikepedia, Baidu Baike, Baidu news and Baidu Tieba. The number of sentences are 21M, 51M, 47M, 54M. respectively. Baidu Baike contains encyclopedia articles written in formal languages, which is used as a strong basis for language modeling. Baidu news provides the latest information about movie names, actor names, football team names, etc. Baidu Tieba is an open discussion forum like Reddits, where each post can be regarded as a dialogue thread. Tieba corpus is used in our DLM task, which will be discussed in the next section.

We perform traditional-to-simplified conversion on the Chinese characters, and upper-to-lower conversion on English letters. We use a shared vocabulary of 17,964 unicode characters for our model.

4.2 DLM

Dialogue data is important for semantic representation, since the corresponding query semantics of the same replies are often similar. ERNIE models the Query-Response dialogue structure on the DLM (Dialogue Language Model) task. As shown in figure 3, our method introduces dialogue embedding to identify the roles in the dialogue, which is different from that of universal sentence encoder (Cer et al., 2018). ERNIE's Dialogue embedding plays the same roles as token type embedding in BERT, except that ERNIE can also represent multi-turn conversations (e.g. QRQ, QRR, QQR, where Q and R stands for "Query" and "Response" respectively). Like MLM in BERT, masks are applied to enforce the model to predict missing words conditioned on both query and response. What's more, we generate fake samples by replacing the query or the response with a randomly selected sentence. The model is designed to judge whether the multi-turn conversation is real or fake.

The DLM task helps ERNIE to learn the implicit relationship in dialogues, which also enhances the model's ability to learn semantic representation. The model architecture of DLM task is compatible with that of the MLM task, thus it is pre-trained alternatively with the MLM task.

4.3 Experiments on Chinese NLP Tasks

ERNIE is applied to 5 Chinese NLP tasks, including natural language inference, semantic similarity, named entity recognition, sentiment analysis, and question answering.

4.3.1 Natural Language Inference

The Cross-lingual Natural Language Inference (XNLI) corpus (Liu et al., 2019) is a crowd-sourced collection for the MultiNLI corpus. The pairs are annotated with textual entailment and translated into 14 languages including Chinese. The labels contains contradiction, neutral and entailment. We follow the Chinese experiments in BERT(Devlin et al., 2018).

4.3.2 Semantic Similarity

The Large-scale Chinese Question Matching Corpus (LCQMC) (Liu et al., 2018) aims at identifying whether two sentences have the same intention. Each pair of sentences in the dataset is associated with a binary label indicating whether the two sentences share the same intention, and the task can be formalized as predicting a binary label.

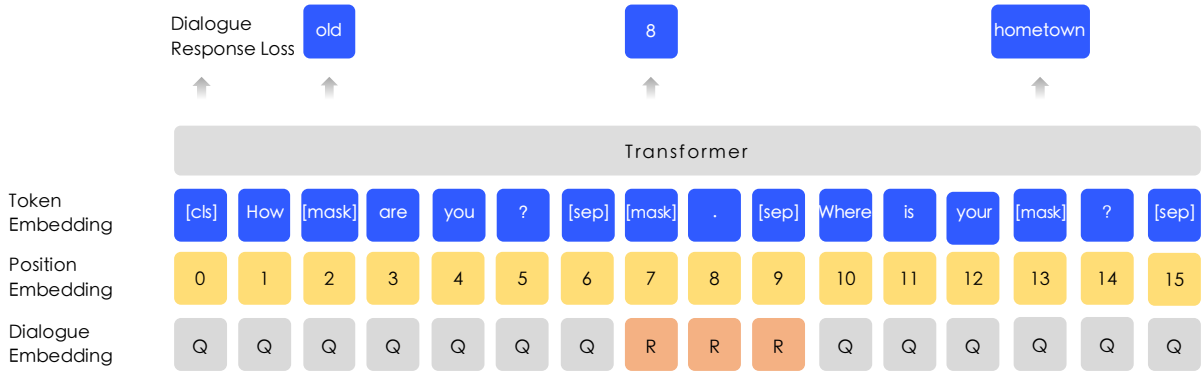

Figure 3: Dialogue Language Model. Source sentence: [cls] How [mask] are you [sep] 8 . [sep] Where is your [mask] ? [sep]. Target sentence (words the predict): old, 8, hometown)

4.3.3 Name Entity Recognition

The MSRA-NER dataset is designed for named entity recognition, which is published by Microsoft Research Asia. The entities contains several types including person name, place name, organization name and so on. This task can be seen as a sequence labeling task.

4.3.4 Sentiment Analysis

ChnSentiCorp (Song-bo) is a dataset which aims at judging the sentiment of a sentence. It includes comments in several domains such as hotels, books and electronic computers. the goal of this task is to judge whether the sentence is positive or negative.

4.3.5 Retrieval Question Answering

The goal of NLPCC-DBQA dataset (`http://tcci.ccf.org.cn/conference/2016/dldoc/evagline2.pdf`) is to select answers of the corresponding questions. The evaluation methods on this dataset include MRR (Voorhees, 2001) and F1 score.

4.4 Experiment results

The test results on 5 Chinese NLP tasks are presented in Table 1. It can be seen that ERNIE outperforms BERT on all tasks, creating new state-of-the-art results on these Chinese NLP tasks. For the XNLI, MSRA-NER, ChnSentiCorp and nlpcc-dbqa tasks, ERNIE obtains more than 1% absolute accuracy improvement over BERT. The gain of ERNIE is attributed to its knowledge integration strategy.

4.5 Ablation Studies

To better understand ERNIE, we perform ablation experiments over every strategy of ERNIE in this section.

4.5.1 Effect of Knowledge Masking Strategies

We sample 10% training data from the whole corpus to verify the effectiveness of the knowledge masking strategy. Results are presented in Table 2. We can see that adding phrase-level mask to the baseline word-level mask can improve the performance of the model. Based on this, we add the entity-level masking strategythe performance of the model is further improved. In addition. The results also show that with 10 times larger size of the pre-training dataset, 0.8% performance gain is achieved on XNLI test set.

4.5.2 Effect of DLM

Ablation study is also performed on the DLM task. we use 10% of all training corpus with different proportions to illustrate the contributions of DLM task on XNLI develop set. we pre-train ERNIE from scratch on these datasets, and report average result on XNLI task from 5 random restart of fine-tuning. Detail experiment setting and develop set result is presented in Table 3, We can see that 0.7%/1.0% of improvement in develop/test accuracy is achieved on this DLM task.

4.6 Cloze Test

To verify ERNIE's knowledge learning ability, We use several Cloze test samples (Taylor, 1953) to examine the model. In the experiment, the name entity is removed from the paragraphs and the model need to infer what it is. Some cases are show in Figure 4. We compared the predictions of BERT and ERNIE.

In case 1, BERT try to copy the name appeared in the context while ERNIE remembers the knowl-

Table 1: Results on 5 major Chinese NLP tasks

Task	Metrics	Bert		ERNIE	
		dev	test	dev	test
XNLI	accuracy	78.1	77.2	79.9 (+1.8)	78.4 (+1.2)
LCQMC	accuracy	88.8	87.0	89.7 (+0.9)	87.4 (+0.4)
MSRA-NER	F1	94.0	92.6	95.0 (+1.0)	93.8 (+1.2)
ChnSentiCorp	accuracy	94.6	94.3	95.2 (+0.6)	95.4 (+1.1)
nlpcc-dbqa	mrr	94.7	94.6	95.0 (+0.3)	95.1 (+0.5)
	F1	80.7	80.8	82.3 (+1.6)	82.7 (+1.9)

Table 2: XNLI performance with different masking strategy and dataset size

pre-train dataset size	mask strategy	dev Accuracy	test Accuracy
10% of all	word-level(chinese character)	77.7%	76.8%
10% of all	word-level&phrase-level	78.3%	77.3%
10% of all	word-level&phrase-leve&entity-level	78.7%	77.6%
all	word-level&phrase-level&entity-level	79.9 %	78.4%

Table 3: XNLI finetuning performance with DLM

corpus proportion(10% of all training data)	dev Accuracy	test Accuracy
Baike(100%)	76.5%	75.9%
Baike(84%) / news(16%)	77.0%	75.8%
Baike(71.2%)/ news(13%)/ forum Dialogue(15.7%)	77.7%	76.8%

No	Text	Predict by ERNIE	Predict by BERT	Answer
1	2006年9月，_____与张柏芝结婚，两人婚后育有两儿子——大儿子Lucas谢振轩，小儿子Quintus谢振南；	谢霆锋	谢振轩	谢霆锋
	In September 2006, _____ married Cecilia Cheung. They had two sons, the older one is Zhenxuan Xie and the younger one is Zhennan Xie.	Tingfeng Xie	Zhenxuan Xie	Tingfeng Xie
2	戊戌变法，又称百日维新，是_____、梁启超等维新派人士通过光绪帝进行的一场资产阶级改良。	康有为	孙世昌	康有为
	The Reform Movement of 1898, also known as the Hundred-Day Reform, was a bourgeois reform carried out by the reformists such as ____ and Qichao Liang through Emperor Guangxu.	Youwei Kang	Shichang Sun	Youwei Kang
3	高血糖则是由于_____分泌缺陷或其生物作用受损，或两者兼有引起。糖尿病时长期存在的高血糖，导致各种组织，特别是眼、肾、心脏、血管、神经的慢性损害、功能障碍。	胰岛素	糖糖内	胰岛素
	Hyperglycemia is caused by defective _____ secretion or impaired biological function, or both. Long-term hyperglycemia in diabetes leads to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves.	Insulin	(Not a word in Chinese)	Insulin
4	澳大利亚是一个高度发达的资本主义国家，首都为_____。作为南半球经济最发达的国家和全球第12大经济体、全球第四大农产品出口国，其也是多种矿产出口量全球第一的国家。	墨尔本	墨悉本	堪培拉
	Australia is a highly developed capitalist country with _____ as its capital. As the most developed country in the Southern Hemisphere, the 12th largest economy in the world and the fourth largest exporter of agricultural products in the world, it is also the world's largest exporter of various minerals.	Melbourne	(Not a city name)	Canberra (the capital of Australia)
5	_____是中国神魔小说的经典之作，达到了古代长篇浪漫主义小说的巅峰，与《三国演义》《水浒传》《红楼梦》并称为中国古典四大名著。	西游记	《小》	西游记
	_____is a classic novel of Chinese gods and demons, which reaching the peak of ancient Romantic novels. It is also known as the four classical works of China with Romance of the Three Kingdoms, Water Margin and Dream of Red Mansions.	The Journey to the West	(Not a word in Chinese)	The Journey to the West
6	相对论是关于时空和引力的理论，主要由_____创立。	爱因斯坦	卡尔斯所	爱因斯坦
	Relativity is a theory about space-time and gravity, which was founded by _____.	Einstein	(Not a word in Chinese)	Einstein

Figure 4: Cloze test

edge about relationship mentioned in the article. In cases 2 and Case 5, BERT can successfully learn the patterns according to the contexts, therefore correctly predicting the named entity type but failing to fill in the slot with the correct entity. on the contrary, ERNIE can fill in the slots with the correct entities. In cases 3, 4, 6, BERT fills in the slots with several characters related to sentences, but it is hard to predict the semantic concept. ERNIE predicts correct entities except case 4. Although ERNIE predicts the wrong entity in Case 4, it can correctly predict the semantic type and fills in the slot with one of an Australian city. In summary, these cases show that ERNIE performs better in context-based knowledge reasoning.

5 Conclusion

In this paper, we presents a novel method to integrate knowledge into pre-training language model. Experiments on 5 Chinese language processing tasks show that our method outperforms BERT over all of these tasks. We also confirmed that both the knowledge integration and pre-training on heterogeneous data enable the model to obtain better language representation.

In future we will integrate other types of knowledge into semantic representation models, such as using syntactic parsing or weak supervised signals from other tasks. In addition We will also validate this idea in other languages.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Daniel Cer, Yinfei Yang, Sheng Yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, and Chris Tar. 2018. Universal sentence encoder.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

TAN Song-bo. Chnsenticorp.

Wilson L Taylor. 1953. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ellen M Voorhees. 2001. Overview of the trec 2001 question answering track. In *TREC*, pages 42–51.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun Hsuan Sung, and Brian Strope. 2018. Learning semantic textual similarity from conversations.