ZS-BERT: Towards Zero-Shot Relation Extraction with Attribute Representation Learning

Chih-Yao Chen
Progam of Data Science
National Taiwan University
Taipei, Taiwan
r09946001@ntu.edu.tw

Cheng-Te Li
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
chengte@ncku.edu.tw

Abstract

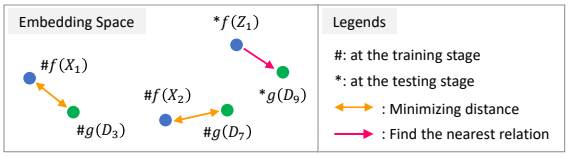
While relation extraction is an essential task in knowledge acquisition and representation, and new-generated relations are common in the real world, less effort is made to predict unseen relations that cannot be observed at the training stage. In this paper, we formulate the zero-shot relation extraction problem by incorporating the text description of seen and unseen relations. We propose a novel multi-task learning model, zero-shot BERT (ZS-BERT), to directly predict unseen relations without hand-crafted attribute labeling and multiple pairwise classifications. Given training instances consisting of input sentences and the descriptions of their relations, ZS-BERT learns two functions that project sentences and relation descriptions into an embedding space by jointly minimizing the distances between them and classifying seen relations. By generating the embeddings of unseen relations and new-coming sentences based on such two functions, we use nearest neighbor search to obtain the prediction of unseen relations. Experiments conducted on two well-known datasets exhibit that ZS-BERT can outperform existing methods by at least 13.54% improvement on F1 score.

1 Introduction

Relation extraction is an important task in the natural language processing field, which aims to infer the semantic relation between a pair of entities within a given sentence. There are many applications based on relation extraction, such as extending knowledge bases (KB) (Lin et al., 2015) and improving question answering task (Xu et al., 2016). Existing approaches to this task usually require large-scale labeled data. However, the labeling cost is a considerable difficulty. Some recent studies generate labeled data based on distant supervision (Mintz et al., 2009; Ji et al., 2017). Nevertheless, when putting the relation extraction task in the wild, existing supervised models cannot well recognize the relations of instances that are extremely rare or even never covered by the training data. That said, in the real-world setting, we should not presume the relations/classes of new-coming sentences are always included in the training data. Thus it is crucial to invent new models to predict *new classes* that are not defined or observed beforehand. Such a task is referred as *zero-shot learning* (ZSL) (Norouzi et al., 2013; Lampert et al., 2014; Ba et al., 2015; Kodirov et al., 2017). The idea of ZSL is to connect seen and the unseen classes by finding an intermediate semantic representation. Unlike the common way to train a supervised model, seen and unseen classes are disjoint at training and testing stages. Hence, ZSL models need to generate transferable knowledge between them. With a model for ZSL relation extraction, we will be allowed to extract unobserved relations, and to deal with new relations resulting from the birth of new entities.

Existing studies on ZSL relation extraction are few and face some challenges. First, while the typical study (Levy et al., 2017) cannot perform zero-shot relation classification without putting more human effort on it, as they solve this problem via pre-defining question templates. However, it is infeasible and impractical to manually create templates of new-coming unseen relations under the zero-shot setting. We would expect a model that can produce accurate zero-shot prediction without the effort of hand-crafted labeling. In this work, we take advantage of the description of relations, which are usually publicly available, to achieve the goal. Second, although there exists studies that also utilize the accessibility of the relation descriptions (Obamuyide and Vlachos, 2018), they simply treat zero-shot prediction as the text entailment task and only output a binary label that indicates whether the entities in the input sentence can be depicted by a given relation description. Such problem formulation requires the impractical execution

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3470–3479
June 6–11, 2021. ©2021 Association for Computational Linguistics



	Input Sentence	Relation	Relation Description
Training	X_1: In 1997, Dennis Crouch and Hester put together a western swing band called "The Time Jumpers"	y_3: **member of** (seen)	D_3: organization, musical group, or club to which the subject belongs
Training	X_2: He had roles in two 2008 films: the sci-fi film "Jumper" and the World War II drama "Defiance"	y_7: **main subject** (seen)	D_7: primary topic of a work
Testing	Z_1: During the Philippine–American War, Mark Twain wrote a short pacifist story titled "The War Prayer"	y_9: **author** (unseen, ground truth)	D_9: main creator(s) of a written work

Figure 1: An example for elaborating our ZS-BERT.

of multiple classifications over all relation descriptions, and cannot make relations comparable with each other.

This paper presents a novel model, *Zero-shot BERT* (**ZS-BERT**), to perform zero-shot learning for relation extraction to cope with the challenges mentioned above. ZS-BERT takes two model inputs. One is the *input sentence* containing the pair of target entities, and the other is the *relation description*, i.e., text describing the relation of two target entities. The model output is the *attribute vector*[1] depicting the relation. The attribute vector can be considered as a semantic representation of the relation, and will be used to generate the final prediction of unseen relations. We think a better utilization of relation descriptions by representation learning is more cost-effective than collecting tons of instances with labeled relations. Therefore, an essential benefit of ZS-BERT is free from heavy-cost crowdsourcing or annotation, i.e., annotating what kind of attribute does a class have, which is commonly used in zero-shot learning problem (Lu et al., 2018; Lampert et al., 2009).

Figure 1 depicts the overview of the proposed ZS-BERT, which consists of five steps. Each training instance is a pair of input sentence X_i and its corresponding relation's description D_j. First, we learn a projection function f that projects the input sentence X_i to its corresponding attribute vector, i.e., sentence embedding. Second, we learn another mapping function g that encodes the relation description D_j as into its corresponding attribute vector, which is the semantic representation of D_j. Third, given the training instance (X_i, D_j), we train ZS-BERT by minimizing the distance between attribute vectors $f(X_i)$ and $g(D_j)$ in the embedding space. Fourth, with the learned $g(D_l)$, we are allowed to project the unseen relation's description D_l into the embedding space so that unseen classes can be as separate as possible for zero-shot prediction. Last, given a new input sentence Z_k, we can use its attributed vector $f(Z_k)$ to find the nearest neighbor in the embedding space as the final prediction. In short, the main idea of ZS-BERT is to learn the representations of relations based on their descriptions, and to align the representations with input sentences, at the training stage. In addition, we exploit the learned alignment projection functions f and g to generate the prediction of unseen relations for the new sentence so that the zero-shot relation extraction can be achieved. Our contributions can be summarized as below.

- Conceptually, we formulate the zero-shot relation extraction problem by leveraging text descriptions of seen and unseen relations. To the best of our knowledge, we are the first attempt to directly predict unseen relation under the zero-shot setting via learning the representations from relation descriptions.

- Technically, we propose a novel deep learning-based model, ZS-BERT[2], to tackle the zero-shot relation extraction task. ZS-BERT learns the projection functions to align the input sentence with its relation in the embedding space, and thus is capable of predicting relations that were not seen during the training stage.

- Empirically, experiments conducted on two well-known datasets exhibit that ZS-BERT can significantly outperform state-of-the-art methods for predicting unseen relations under the ZSL setting. We also show that ZS-BERT can be quickly adapted and generalized to few-shot learning when a small fraction of labeled data for unseen relations is available.

2 Related Work

BERT-based Relation Extraction. Contextual representation of words is effective for NLP tasks. BERT (Devlin et al., 2019) is a pre-training language model that learns useful contextual word representations. BERT can be moderately adopted

[1]The terms, "attribute vector", "embedding", and "representation", are used interchangeably throughout this paper.

[2]Code and implementation details can be accessed via: https://github.com/dinobby/ZS-BERT.

for supervised or few-shot relation extraction. R-BERT (Wu and He, 2019) utilize BERT to generate contextualized word representation, along with entities' information to perform supervised relation extraction and have shown promising result. BERT-PAIR (Gao et al., 2019) makes use of the pre-train BERT sentence classification model for few-shot relation extraction. By pairing each query sentence with all sentences in the support set, they can get the similarity between sentences by pre-trained BERT, and accordingly classify new classes with a handful of instances. These models aim to solve the general relation extraction task, which are more or less having ground truth, rather than having it under the zero-shot setting.

Zero-shot Relation Extraction. Relevant studies on zero-shot relation extraction are limited. To the best of our knowledge, there are two most similar papers, which consider zero-shot relation extraction as two different tasks. Levy et al. (2017) treat zero-shot relation extraction as a question answering task. They manually define 10 question templates to represent relations, and generate the prediction by training a reading comprehension model to answer which relation satisfies the given sentence and question. However, it is required to have human efforts on defining question templates for unseen relations so that ZSL can be performed. Such annotation by domain knowledge is unfeasible in the wild when more unseen relations come. On the contrary, the data requirement of ZS-BERT is relatively lightweight. For each relation, we only need one description that could express the semantic meaning. The descriptions of relations are easier to be collected as we may access them from open resources. Under such circumstances, we may be free from putting additonal effort to the annotation.

Obamuyide and Vlachos (2018) formulate ZSL relation extraction as a textual entailment task, which requires the model to predict whether the input sentence containing two entities matches the description of a given relation. They use Enhanced Sequential Inference Model (ESIM) (Chen et al., 2016) and Conditioned Inference Model (CIM) (Rocktäschel et al., 2015) as their entailment methods. By pairing each input sentence with every relation description, they train the models to answer whether the paired texts are contradiction or entailment. This allow the model to inference on input sentence and unseen relation description pair, thus is able to predict unseen relation accordingly.

3 Problem Definition

Let $Y_s = \{y_s^1, ..., y_s^n\}$ and $Y_u = \{y_u^1, ..., y_u^m\}$ denote the sets of seen and unseen relation labels, respectively, in which $n = |Y_s|$ and $m = |Y_u|$ are the numbers of relations in two sets. Such two sets are disjoint, i.e., $Y_s \cap Y_u = \emptyset$. For each relation label in seen and unseen sets, we denote the corresponding attribute vector as $a_s^i \in \mathbb{R}^{n \times d}$ and $a_u^i \in \mathbb{R}^{m \times d}$, respectively. Given the training set with N samples, consisting of input sentence X_i, entities e_{i1} and e_{i2}, and the description D_i of the corresponding seen relation y_s^j, denoted as $\{S_i = (X_i, e_{i1}, e_{i2}, D_i, y_s^j)\}_{i=1}^N$. Our goal is to train a zero-shot relation extraction model $\mathcal{M}$, i.e., $\mathcal{M}(S_i) \rightarrow y_s^i \in Y_s$, based on the training set such that using $\mathcal{M}$ to predict the *unseen* relation y_u^k of a testing instance S', i.e., $\mathcal{M}(S') \rightarrow y_u^j \in Y_u$, can achieve as better as possible performance.

We train the model $\mathcal{M}$ so that the semantics between input sentence and relation description can be aligned. We learn $\mathcal{M}$ by minimizing the distance between two embedding vectors $f(X_i)$ and $g(D_i)$, where learnable functions f and g project X_i and D_i into the embedding space, respectively. When new unseen relation y_u^j and its description is in hand, we can project the description of y_u^j to the embedding space by function g. When testing, new instance $S' = (Z_j, e_{j1}, e_{j2}, D_j)$ is input, in which Z_i denotes new sentence containing entities e_{j1} and e_{j2}, we project Z_i to the embedding space by our learned function f, and find the nearest neighboring unseen relation y_u^j, where Z_i and y_u^i are both unknown at the training stage.

4 The Proposed ZS-BERT Model

We give an overview of our ZS-BERT in Figure 2. The input sentence X_i is tokenized and sent into the upper-part ZS-BERT encoder to obtain contextual representation. We specifically extract the representation of [CLS], H_0, and two entities' representations H_e^1, H_e^2, and then concatenate them to derive sentence embeddings $\hat{a}_s^i$, by a fully-connected layer and activation operation. In the bottom part, we use Sentence-BERT (Reimers and Gurevych, 2019) to obtain attribute vector a_s^i for seen relations by encoding the corresponding description of relation D_i. We train ZS-BERT under a multi-task learning structure. One task is to minimize the distance between attribute vector a_s^i and sentence embedding $\hat{a}_s^i$. The other is to classify the seen relation y_s^j at the training stage, in which a softmax

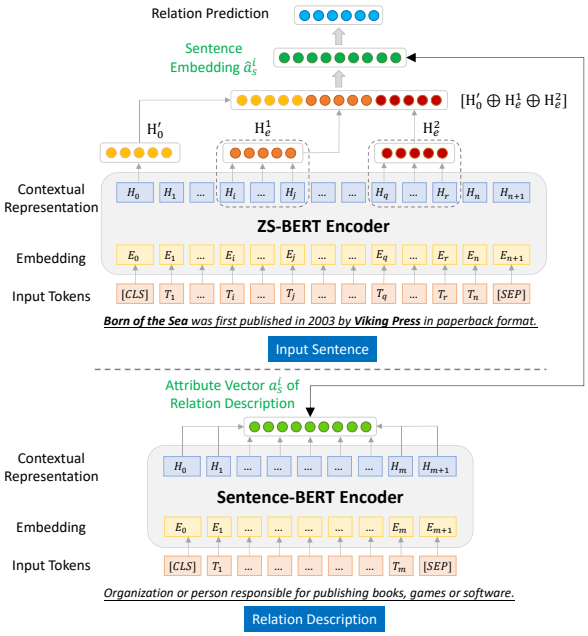


Figure 2: The overall architecture of our model.

layer that accepts relation embedding is used to produce the relation classification probability. At the testing stage, by obtaining the embeddings of new-coming sentences and unseen relations, we use $\hat{a}_s^i$ and nearest neighbor search to obtain the prediction of unseen relations.

4.1 Learning Relation Attribute Vectors

For each seen and unseen relation, we learn its representation that depicts the corresponding semantic attributes based on relation description D_i. Most relations are well-defined and their descriptions are accessible from online open resources such as Wikidata [3]. We feed relation description D_i into a pre-trained Sentence-BERT encoder (Reimers and Gurevych, 2019) to generate the sentence-level representation as the attribute vector a^i of relations. This procedure is shown in the bottom part of Figure 2. The ground truth relation of the example is *publisher*, along with its description *Organization or person responsible for publishing books, games or software.* We feed only the relation description to the Sentence-BERT in order to get the attribute vector. That said, we consider the derived Sentence-BERT to be a projection function g that transforms the relation description D_i into a^i. Note that the relation attribute vectors produced by Sentence-BERT are fixed during model training.

[3] https://www.wikidata.org

4.2 Input Sentence Encoder

We utilize BERT (Devlin et al., 2019) to generate the contextual representation of each token. We first tokenize the input sentences X_i with WordPiece tokenization (Sennrich et al., 2016). Two special tokens [CLS] and [SEP] are appended to the first and last positions, respectively. Since the entity itself does matter in relation extraction, we use an entity marker vector, consisting of all zeros except the indices that entities appear in a sentence, to indicate the positions of entities e_{i1} and e_{i2}. Let H_0 be the hidden state of the first special token [CLS]. We use a $tanh$ activation function, together with a fully connected layer, to derive the representation vector H'_0, given by: $H'_0 = W_0[tanh(H_0)] + b_0$, where W_0 and b_0 are learnable parameters for weights and biases. We obtain the hidden state vectors of two entities, H^1_e and H^2_e, by averaging their respective tokens' hidden state vectors. The entity can be recognized via simple element-wise multiplication between entity marker vector and token hidden vector. Specifically, if an entity e consists of multiple tokens and the indices range from q to r, we average the hidden state vectors, and also add an activation operation with a fully connected layer to generate its representation of that entity, given by: $H^c_e = W_e\left[tanh\left(\frac{1}{r-q+1}\sum_{t=q}^{r} H_t\right)\right] + b_e$, where $c = 1, 2$. Note that the representations of two entities $H^c_e (c = 1, 2)$ in the sentence shares the same parameters W_e and b_e. Then we learn the attribute vector $\hat{a}_s^i$ by concatenating H'_0, H^1_e, and H^2_e, followed by a hidden layer, given by:

$$\hat{a}_s^i = W_1(tanh([H'_0 \oplus H^1_e \oplus H^2_e])) + b_1, \quad (1)$$

where W_1 and b_1 are learnable parameters , the dimensionality of $\hat{a}_s^i$ is d, and $\oplus$ is the concatenation operator.

4.3 Model Training

The training of our ZS-BERT model consists of two objectives. The first is to minimize the distance between input sentence embedding a_s^i and the corresponding relation attribute vector $\hat{a}_s^i$ (i.e., *positive* pairs), meanwhile to ensure embedding pairs between input sentence embedding and mismatched relation (i.e., *negative* pairs) to be farther away from each other. The black arrow connecting a_s^i and $\hat{a}_s^i$ in Figure 2 is a visualization to indicate that we take both a_s^i and $\hat{a}_s^i$ into consideration to achieve this goal. This is also reflected in the first term of our proposed loss function introduced

Table 1: Datasets. "avg. len." is average sentence len.

	#instances	#entities	#relations	avg. len.
Wiki-KB	1,518,444	306,720	354	23.82
Wiki-ZSL	94,383	77,623	113	24.85
FewRel	56,000	72,954	80	24.95

below. The second objective is to maximize the accuracy of relation classification based on seen relations using cross entropy loss. We transform the relation embedding, along with a softmax layer, to generate a n-dimensional ($n = |Y_s|$) classification probability distribution over seen relations: $p(y_s|X_i, \theta) = softmax(W^*(tanh(\hat{a}_s^i)) + b^*)$, where $y_s \in Y_s$ is the seen relation, θ is the model parameter, $W^* \in \mathbb{R}^{n \times h}$, h is the dimension of hidden layer, and $b^* \in \mathbb{R}^n$. Note that we do not use the probability distribution but the input sentence embedding $\hat{a}_s^i$ produced *intermediately* for predicting unseen relations under zero-shot settings.

The objective function of ZS-BERT is as follows:

$$L = (1-\alpha)\sum_{i}^{N} \max(0, \gamma - a_s^i \cdot \hat{a}_s^i + \max_{i \neq j}(a_s^i \cdot \hat{a}_s^j)) - \alpha \sum_{i}^{N} y_s^i log(\hat{y_s^i}), \quad (2)$$

where N is the number of samples, a_s^i is the relation attribute vector, and $\hat{a}_s^i$ is the input sentence embedding. The first term in Eq. (2) sets a margin $\gamma > 0$ such that the inner product of the positive pair (i.e., $a_s^i \cdot \hat{a}_s^i$) must be higher than the maximum of the negative one (i.e., $\max_{i \neq j}(a_s^i \cdot \hat{a}_s^j)$) for more than a pre-decided threshold γ. With the introduction of γ, the loss will be increased owing to the difference between the positive and the closest negative pairs. This design of loss function can be viewed as ranking the correct relation attribute higher than the closest incorrect one. In addition, γ is also utilized to avoid the embedding space from collapsing. If we consider only minimizing the distance of positive pair using loss like Mean Squared Error, the optimization may lead to the result that every vector in the embedding space is too close to one another. We will examine how different γ values affect the performance in the experiment. To maintain low computational complexity, we consider only those mismatched relations within a batch as the negative samples j. The second term in Eq. (2) is a commonly used cross entropy loss, which decreases as the prediction $\hat{y}_s^i$ is correctly classified. Such a multi-task structure is expected to refine the input sentence embeddings and simultaneously bring high prediction accuracy of seen relations.

4.4 Generating Zero-Shot Prediction

With the trained model, when the descriptions of new relations are in hand, we can generate their attribute vectors a_u^j. As the new input sentence Z_i arrives, we can also produce its sentence embedding $\hat{a}_u^i$ via: $\hat{a}_u^i = W_1(tanh([H_0' \oplus H_e^1 \oplus H_e^2])) + b_1$, where W_1 and b_1 are learned parameters. The prediction on unseen relations can be achieved by the nearest neighbor search. For the input sentence embedding $\hat{a}_u^i$, we find the nearest attribute vector a_u^j and consider the corresponding relation as the predicted unseen relation. This can be depicted by: $C(Z_i) = \text{argmin}_j\, dist(\hat{a}_u^i, a_u^j)$, where function C returns the predicted relation of new input sentence Z_i, a_u^j is the j-th attribute vector among all unseen relations in the embedding space, $\hat{a}_u^i$ is the new input sentence embedding, and $dist$ is a distance computing function. Here negative inner product is used as $dist$ since we aim to consider the nearest neighboring relation as the predicted outcome.

5 Experiments

5.1 Evaluation Settings

Datasets. Two datasets are employed, **Wiki-ZSL** and **FewRel** (Han et al., 2018). Wiki-ZSL is originated from Wiki-KB (Sorokin and Gurevych, 2017), and is generated with distant supervision. That said, in Wiki-ZSL, *entities* are extracted from complete articles in Wikipedia, and are linked to the Wikidata knowledge base so that their *relations* can be obtained. Since $395,976$ instances (about 26% of the total data) do not contain relations in the original Wiki-KB data, we neglect instances with relation "none". To ensure having sufficient data instances for each relation in zero-shot learning, we further filter out the relations that appear fewer than 300 times. Eventually, we can have yields **Wiki-ZSL**, a subset of Wiki-KB.

On the other hand, FewRel (Han et al., 2018) is compiled by a similar way to collect entity-relation triplet with sentences, but had been further filtered by crowd workers. This ensures the data quality and class balance. Although FewRel is originally proposed for few-shot learning, it is also suitable for zero-shot learning as long as the relation labels within training and testing data are disjoint. The statistics of Wiki-KB, Wiki-ZSL and FewRel datasets are shown in Table 1.

ZSL Settings. We randomly select m relations as *unseen* ones ($m = |Y_u|$), and randomly split the whole dataset into training and testing data, meanwhile ensuring that these m relations do not appear in training data so that $Y_s \cap Y_u = \emptyset$. We repeat the experiment 5 times for random selection of m relations and random training-testing splitting, and report the average results. We will also vary m to examine how performance is affected. We use *Precision* (P), *Recall* (R), and *F1* as the evaluation metrics. As for the hyperparameters and configuration of ZS-BERT, we use Adam (Kingma and Ba, 2014) as the optimizer, in which the initial learning rate is $5e-6$, the hidden layer size is 768, the dimension of input sentence embedding and attribute vector is 1024, the batch size is 4, $\gamma = 7.5$, and $\alpha = 0.4$.

Competing Methods. The compared methods consist of two categories, supervised relation extraction (SRE) models and text entailment models. The former includes CNN-based SRE (Zeng et al., 2014), Bi-LSTM SRE (Zhang et al., 2015), Attentional Bi-LSTM SRE (Zhou et al., 2016), and R-BERT (Wu and He, 2019). These SRE models use different ways to extract features from the input sentences and perform prediction. They have achieved great performance with fully supervision but fail to carry out zero-shot prediction. To make them capable of zero-shot prediction, also to have fair comparison, instead of originally using a softmax layer to output a probability vector whose dimension is equal to the seen relations, we change the last hidden layer of each SRE competing method to a fully-connected layer with a $tanh$ activation function, and the embedding dimension d is the same as ZS-BERT. The nearest neighbor search is applied over input sentence embeddings and relation attribute vectors to generate zero-shot prediction.

Two text entailment models, ESIM (Chen et al., 2016) and CIM (Rocktäschel et al., 2015), are also used for comparison. These two models follow a well-known implementation (Obamuyide and Vlachos, 2018) that formulates zero-shot relation extraction as a text entailment task, which accepts sentence and relation description as input, and output a binary label indicating whether they are semantically matched. ESIM uses bi-LSTM (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) to encode two input sequences, passes them through the local inference model, and produces the prediction via a softmax layer. CIM replaces the bi-LSTM block with a conditional version, i.e., the representation of sentence is conditioned on its relation description. Note that although there exist other zero-shot relation extraction approaches such as the approach proposed by Levy et al. (2017), their approach to formulate the ZSL task and their data requirement are quite different with our present work. To be specific, their method requires pre-defined question template, whereas our model does not. Hence it would be unfair to compare with those approaches.

Table 2: Results with different m values in percentage.

	Wiki-ZSL			FewRel		
	m=5			m=5		
	P	R	F1	P	R	F1
CNN	30.31	32.17	30.92	36.41	38.69	37.42
Bi-LSTM	36.73	40.44	38.62	41.99	50.25	45.66
Att Bi-LSTM	35.58	41.26	38.21	39.52	47.24	42.95
R-BERT	39.22	43.27	41.15	42.19	48.61	45.17
ESIM	48.58	47.74	48.16	56.27	58.44	57.33
CIM	49.63	48.81	49.22	58.05	61.92	59.92
ZS-BERT	**71.54**	**72.39**	**71.96**	**76.96**	**78.86**	**77.90**
	m=10			m=10		
	P	R	F1	P	R	F1
CNN	20.86	23.61	22.08	22.37	28.15	24.85
Bi-LSTM	25.33	27.91	26.56	24.52	32.02	27.77
Att Bi-LSTM	24.98	29.13	26.90	24.24	31.32	27.28
R-BERT	26.18	29.69	27.82	25.52	33.02	28.20
ESIM	44.12	45.46	44.78	42.89	44.17	43.52
CIM	46.54	47.90	45.57	47.39	49.11	48.23
ZS-BERT	**60.51**	**60.98**	**60.74**	**56.92**	**57.59**	**57.25**
	m=15			m=15		
	P	R	F1	P	R	F1
CNN	14.58	17.68	15.92	14.17	20.26	16.67
Bi-LSTM	16.25	18.94	17.49	16.83	27.62	20.92
Att Bi-LSTM	16.93	18.54	17.70	16.48	26.36	20.28
R-BERT	17.31	18.82	18.03	16.95	19.37	18.08
ESIM	27.31	29.62	28.42	29.15	31.59	30.32
CIM	29.17	30.58	29.86	31.83	33.06	32.43
ZS-BERT	**34.12**	**34.38**	**34.25**	**35.54**	**38.19**	**36.82**

5.2 Experimental Results

Main Results. The experiment results by varying m unseen relations are shown in Table 2. First, it can be apparently found that the proposed ZS-BERT steadily outperforms the competing methods over two datasets when targeting at different numbers of unseen relations. The superiority of ZS-BERT gets more significant on $m = 5$. Such results

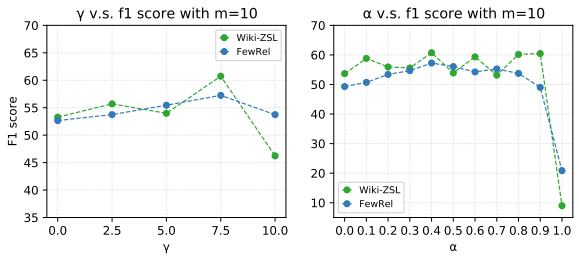


Figure 3: Effects on varying the margin parameter γ and balance coefficient α with m=10 on both datasets.

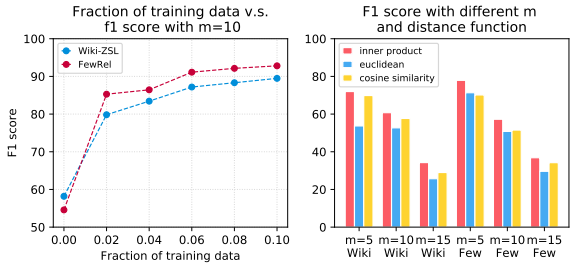


Figure 4: Left: Results of ZS-BERT with different fractions of unseen instances availble for training, in which 0.0 refers to the zero-shot result. Right: Results on different distance functions and varied m.

not only validate the effectiveness of leveraging relation descriptions, but also prove the usefulness of the proposed multi-task learning structure that better encodes the semantics of input sentences and have relation attribute vectors been differentiated from each other. Second, although the text entailment models ESIM and CIM perform well among competing methods, their performance is still obviously lower than ZS-BERT. The reason is that their approaches cannot precisely distinguish the semantics of input sentences and relation descriptions in the embedding space. Third, we also find that the improvement of ZS-BERT gets larger when m is smaller. Increasing m weakens the superiority of ZS-BERT. It is straightforward that as the number of unseen relations increases, it becomes more difficult to predict the right relation since the possible choices have increased. We also speculate another underlying reason is that although ZS-BERT can effectively capture the latent attributes for each relation, relations themselves could be to some extent semantically similar to one another, and more unseen relations will increase the possibility that obtains a predicted relation that is semantically close but actually wrong. To verify this conjecture, we will give an example in the case study.

Hyperparameter Sensitivity. We examine how primary hyperparameters, including the value of margin parameter γ and the balance coefficient α in Eq. 2, affect the performance of ZS-BERT. By fixing $m = 10$ and varying γ and α, the results in terms of F1 scores on two datasets are exhibited on Figure 3. It is noteworthy that γ does have an impact on performance, since it brings the condition on whether to increase the loss value, which is determined by the difference between the positive pair and negative pair. Nevertheless, not always the higher values of γ lead to better performance. This is reasonable that when γ is too low, the distance between the positive pair and negative pair would not be far enough. Thus, when performing nearest neighbor search, it is more likely to reach the wrong relations. In contrast, when γ gets too high, it is hard for the training process to converge at the point that the distance between relations is expected to be that high. We would suggest setting $\gamma = 7.5$ to derive satisfying results across datasets. As for the balance coefficient α in the loss function, we find that $\alpha = 0.4$ can achieve the best performance, indicating that the margin loss plays a more significant role in training ZS-BERT. Also notice that when $\alpha = 1.0$, the performance drops dramatically, showing that the margin loss is essential to our model. This is also reasonable that since our model relies on the quality of embeddings, therefore totally relying on cross entropy loss leads to failure of zero-shot prediction. The better separation between embeddings of different relations, the more likely our model can generate the accurate zero-shot prediction. In addition, while the nearest neighbor search is performed to generate the zero-shot prediction, we think the choice of distance computing function $dist()$ can also be an hyperparameter. By applying inner product, Euclidean distance, and the cosine similarity as $dist()$ in ZS-BERT, we report their F1 scores with different m on two datasets in the right of Figure 4. The results inform us that inner product is a proper distance function for zero-shot relation extraction with ZS-BERT.

Few-shot Prediction. To understand the capability of ZS-BERT, we conduct the experiment of few-shot prediction. By following the setting of an existing work (Obamuyide and Vlachos, 2018), we make a small fraction of unseen data instances available at the training stage. That said, for each originally unseen relation, we move a small fraction of its sentences, along with the relation de-

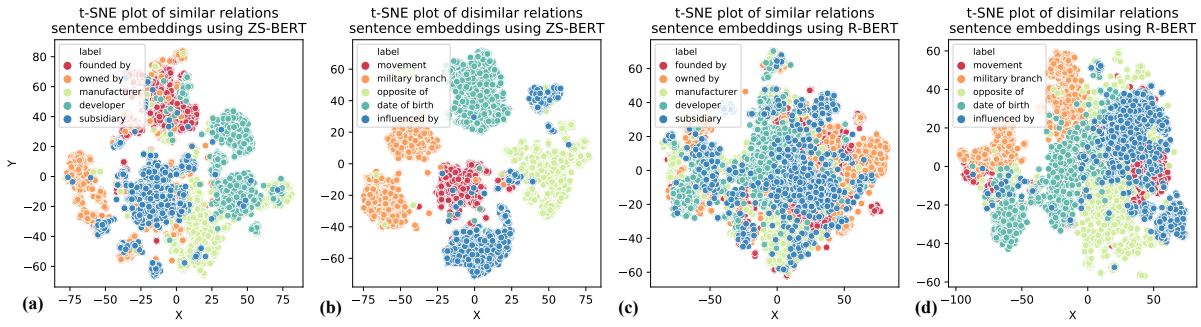


Figure 5: t-SNE visualization of the sentence embeddings for similar (a)(c) and dissimilar relations (b)(d).

Table 3: List of four cases, in which head and tail entities are highlighted by green and blue, respectively.

	Input Sentence	True	Predicted
(1)	When promoting **Anaconda**, **Minaj** confirmed plans of a tour in support of **The Pinkprint** in an interview with Carson Daly on AMP Radio.	tracklist	publisher
(2)	**Heaven** and **Hell** as they are understood in Christian theology are roughly analogous to the Jewish Olam habah and Gehenna , with certain major differences.	opposite of	influenced by
(3)	(TOEO) was an MMORPG set in the world of the popular **Namco** PlayStation title, **Tales of Eternia**.	publisher	manufacturer
(4)	The inverse of **admittance** is **impedance**, and the real part of admittance is conductance.	opposite of	influenced by

scription, from the testing to the training stage. By varying the fraction in x-axis, we report the results of few-shot prediction in Figure 4. We can find that that ZS-BERT can reach about 80% on F1 score with only 2% of unseen instances as supervision. Such results demonstrate the ability to recognize rare samples and the capability of few-shot learning for the proposed ZS-BERT. As expected, the more instances belonging to unseen relations available at the training stage, the higher the F1 score is. When the fraction equals to 10%, ZS-BERT can even achieve 90% F1 score on Wiki-ZSL dataset.

5.3 Case Study

We categorize four types of incorrectly predicted unseen relations for the analysis: (1) The predicted relation is not precise for the targeted entity pair but may be suitable for other entities that also appear in the sentence. (2) The true relation is not appropriate because it comes from distant supervision. (3) The predicted relation is ambiguous or is a synonym of other relations. (4) The relation is wrongly predicted but should be able to be correctly classified. For each of these four types, we provide an example listed in Table 3. In case (1), the targeted entities are **Anaconda** and **The Pinkprint**, and ZS-BERT yields *publisher* as the prediction, which is actually correct if the targeted entities are **Anaconda** and **Minaj**. This shows ZS-BERT is able to infer the possible relation for entities in the given sentence, but sometimes could be misled by non-targeted entities even though we have an entity mask to indicate the targeted entities. In case (2), it shows the noise originated from distant labeling. That is, even human being cannot identify the relation between **Heaven** and **Hell** is *opposite of* in this specific sentence. They just happened to appear together and their relation recorded in Wikidata is *opposite of*. In case (3), the predicted unseen relation is *manufacturer*, while the ground truth is *publisher*. Both *manufacturer* and *publisher* describe someone make or produce something, although their domains are slightly different. This exhibits the capability of ZS-BERT to identify the input sentence with an *abstract* attribute because relations possessing similar semantics will have similar attribute vectors in the embedding space. Finally, in case (4), the model gives a wrong prediction that is not even close or related, which may due to the noise or information loss when transferring knowledge between relations.

Among these four groups, we are especially interested in case (3) since the semantic similarity between relations in the embedding space greatly impacts the performance. We select five semantically-distant relations, and the other five relations that possess similar semantics between two or three of them, to inspect their distributions in the embedding space. We feed sentences with these relations and generate their embeddings using ZS-BERT and R-BERT (Wu and He, 2019) for comparison. We choose R-BERT because it is the strongest embedding-based competing method for zero-shot prediction by nearest neighbor search. Note that since the predictions by text entailment-based models, ESIM and CIM, neither resort to similarity search nor directly predict unseen relation at one time, we cannot have them compared in this analysis. We visualize the embedding space by

t-SNE (Maaten and Hinton, 2008), as shown in Figure 5. We can find that when the relations are somewhat similar in their meanings (Figure 5(a),(c)), some of the data points are mingled with different clusters, as they indeed have close semantic relationships. Take *subsidiary* and *owned by* as examples, *Company A is a subsidiary of company B* and *Company A is owned by company B* refer to the same thing. This happens on both ZS-BERT and R-BERT but to a different extent. It is obvious that the embeddings produced by R-BERT are more tangled. We also plot the other five relations that there is no ambiguity among them (Figure 5(b),(d)). Apparently their embeddings are more separated between different relations. It is also obvious that the embeddings generated by ZS-BERT lead to larger inter-relation distance. This again exhibits the usefulness of the proposed ranking loss and multi-task learning structure.

6 Conclusions

In this work, we present a novel and effective model, ZS-BERT, to tackle the zero-shot relation extraction task. With the multi-task learning structure and the quality of contextual representation learning, ZS-BERT can not only well embed input sentences to the embedding space, but also substantially improve the performance. We have also conducted extensive experiments to study different aspects of ZS-BERT, from hyperparameter sensitivity to case study, and eventually show that ZS-BERT can steadily outperform existing relation extraction models under zero-shot settings. Furthermore, learning effective embeddings for relations might also be helpful to semi-supervised learning or few-shot learning by utilizing prototypes of relations as the auxiliary information.

Acknowledgements

This work is supported by Ministry of Science and Technology (MOST) of Taiwan under grants 109-2636-E-006-017 (MOST Young Scholar Fellowship) and 109-2221-E-006-173, and also by Academia Sinica under grant AS-TP-107-M05.

References

J. L. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4247–4255.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

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.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3060–3066. AAAI Press.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

E. Kodirov, T. Xiang, and S. Gong. 2017. Semantic autoencoder for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4447–4456.

C. H. Lampert, H. Nickisch, and S. Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958.

C. H. Lampert, H. Nickisch, and S. Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Jiang Lu, Jin Li, Ziang Yan, Fenghua Mei, and Changshui Zhang. 2018. Attribute-based synthetic network (abs-net): Learning more from pseudo feature representations. *Pattern Recognition*, 80:129–142.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. 2013. Zero-shot learning by convex combination of semantic embeddings.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 2361–2364, New York, NY, USA. Association for Computing Machinery.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on Freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336, Berlin, Germany. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.