Prompt-based Zero-shot Relation Classification with Semantic Knowledge Augmentation

Anonymous ACL submission

Abstract

In relation classification, recognizing unseen (new) relations for which there are no training instances is a challenging task. We propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to recognize unseen relations under the zero-shot setting. We present a new word-level sentence translation rule and generate augmented instances with unseen relations from instances with seen relations using that new rule. We design prompts based on an external knowledge graph to integrate semantic knowledge information learned from seen relations. Instead of using the actual label sets in the prompt template, we construct weighted virtual label words. We learn the representations of both seen and unseen relations with augmented instances and prompts. We then calculate the distance between the generated representations using prototypical networks to predict unseen relations. Extensive experiments conducted on three public datasets show that ZS-SKA outperforms state-of-the-art methods under the zero-shot scenarios. Our experimental results also demonstrate the effectiveness and robustness of ZS-SKA.

1 Introduction

Relation classification aims to infer the semantic relation between a pair of entities in a sentence. However, existing approaches based on supervised learning (Zhu et al., 2019; Li and Tian, 2020) or few-shot learning (Gao et al., 2019; Ren et al., 2020; Dong et al., 2020) still require labeled data. They can not catch up with a dynamic and open environment where new classes emerge. In the real-world setting, the classes of instances are sometimes rare or never seen in the training data. Thus, we tend to learn a model similar to the way humans learn and recognize new concepts. Such a task is referred to as zero-shot learning (ZSL). We follow the definition of a more generalized zero-shot setting that partial classes are new (Wenpeng Yin and Roth, 2019).

Zero-shot relation classification aims to classify relations of name entities in a sentence that are absent from the learning stage. Existing approaches on zero-shot relation classification still have limitations. First, some models perform zero-shot relation classification by listing questions that define the relation’s slot values (Levy et al., 2017). These models have a strong assumption that an excellent question-answering model is learned, and all values extracted from this model are correct. This is impractical in the real-world setting. Second, some existing studies formulate relation extraction as a text entailment task (Obamuyide and Vlachos, 2018). They only predict a binary label indicating whether the name entities in the given sentence can be described by a given description. Third, some state-of-the-art models leverage side (auxiliary) information to tackle zero-shot tasks. They focus on class names/descriptions semantic information, losing the connection or relationships between seen relations and unseen relations (Gong and Eldardiry, 2021; Chen and Li, 2021).

To address the above challenges, we propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to perform zero-shot learning for relation classification. We first implement data augmentation by a word-level sentence translation to generate augmented instances with unseen relations from training instances with seen relations. The super-class of the triplet (subject, relation, object) for augmented instances is the same as the triplet of training instances. We follow a new generation rule introduced in Sec. 3.2.1 to generate high-quality augmented instances for training in zero-shot settings. Note that ZS-SKA is trained only on labeled data from seen classes and augmented data generated from seen classes.

Secondly, inspired by prompt-tuning on pre-trained language models (Schick and Schütze, 2021a,b), we design the prompts based on a knowledge graph to integrate semantic knowledge to gen-
erally infer the features of unseen relations using patterns learned from seen relations. For prompt design, we consider semantic knowledge information, including relation descriptions, super-class of relations and name entities, and a general knowledge graph to effectively learn the unseen relations. Instead of using the real label word directly in the prompt template, we automatically search a set of appropriate label words based on the knowledge graph for each label. The weight of each appropriate label word is calculated based on its semantic knowledge information in Sec. 3.2.2. We calculate the distance between each appropriate label with the true label itself to help denoise the set of appropriate label words. Then, we construct virtual label words in the prompt by weighted averaging all appropriate label word candidates.

Finally, we apply prototypical networks (Snell et al., 2017) to compute a prototype representing each relation. Each prototype is the mean vector of embedded (augmented) sentences with prompts belonging to one relation. Euclidean distance is calculated between query sentence embeddings with prototypes to classify relations. The contributions of this paper can be summarized as follows:

- We propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to predict unseen relations under the zero-shot setting. Unlike some previous works, ZS-SKA considers semantic information from different granularities and does not rely on other complex models.
- We present a new word-level sentence translation rule to generate augmented instances with unseen relations from instances with seen relations. The augmented sentences are then used as the training instances for unseen relations.
- We propose prompts based on an external knowledge graph to integrate semantic knowledge learned from seen relations. We construct weighted virtual label words for mask in prompt template instead of using actual label sets.
- We demonstrate that ZS-SKA significantly outperforms state-of-the-art methods for predicting unseen relations under the ZSL setting on three public datasets. Results show the effectiveness and robustness of ZS-SKA.

2 Related Work

2.1 Prompt Learning in NLP

With the development of Generative Pre-trained Transformer 3 (GPT-3) (Brown et al., 2020), prompt-based learning has received considerable attention. Language prompts have been proved to be effective in downstream tasks leveraging pre-trained language models (Trinh and Le, 2018; Davison et al., 2019; Petroni et al., 2019). Human-designed prompts have achieved promising results in few-shot learning for sentiment classification (Schick and Schütze, 2021a,b). To avoid labor-intensive prompt design, studies explore prompts that are generated automatically (Shin et al., 2020; Jiang et al., 2020; Gao et al., 2021). However, most of the studies focus on supervised or few-shot learning on text classification (Hu et al., 2021; Han et al., 2021; Gu et al., 2021), relation classification (Han et al., 2021; Chen et al., 2021b) and name entity recognition (Ma et al., 2021).

2.2 Zero-shot Relation Classification

Relation extraction is the problem of extracting semantic relations between two name entities within a given sentence. When no training instances are available, some studies use zero-shot relation classification to extract unseen relation types. This is typically done using question-answering models. In particular, by listing questions that define the relation’s slot values (Levy et al., 2017; Cetoli, 2020). To avoid relying on question-answering models, some studies formulate relation extraction as a text entailment task and utilize the accessibility of the relation descriptions (Obamuyide and Vlachos, 2018; Qin et al., 2020; Gong and Eldardiry, 2021; Chen and Li, 2021). However, these models only utilize class names semantic information, losing the connections between relations. Other studies focus on establishing the connection between relations with knowledge graph (Li et al., 2020). Nevertheless, they miss the semantic information from name entities. Inspired by data augmentation from knowledge graph in text classification tasks (Zhang et al., 2019; Chen et al., 2021a) and prompt-based few-shot learning (Hu et al., 2021), we propose a prompt-based zero-shot relation classification framework incorporating external knowledge from the knowledge graph.
3 Methodology

In this section, we introduce the overall framework as shown in Figure 1 of our proposed prompt-based ZS-SKA, starting with problem formulation.

3.1 Problem Definition

We follow the definition of zero-shot from (Wen-peng Yin and Roth, 2019) and the same settings from (Chen and Li, 2021; Gong and Eldardiry, 2021) to conduct our experiments. This is a more generalized zero-shot setting that partial labels are unseen. Given labeled instances belonging to a set of seen classes \( S \), a model \( M : X \rightarrow Y \) is learned, where \( Y = S \cup U \); \( U \) is the unseen class.

For relation classification task, let \( R_s = \{r^1_s, \ldots, r^m_s\} \) and \( R_u = \{r^1_u, \ldots, r^n_u\} \) denote the sets of seen and unseen relations, where \( m = |R_s| \) and \( n = |R_u| \) are the number of relations in the two disjoint sets, i.e., \( R_s \cap R_u = \emptyset \). Given the training set consisting of seen relations \( R_s \) with their corresponding input sentences \( X_i \) and name entities \( e_{i1} \) and \( e_{i2} \), unseen relations \( R_u \), super-class of name entities \( S(e_{i1}), S(e_{i2}) \), relates to \( R_u \) and external knowledge graph \( G \). Our goal is to train a zero-shot relation classification model \( M \) to learn the representations of both seen and unseen relations. \( M \) is learned by minimizing the semantic distance between the embedding of the input and relation representations built from the knowledge graph.

3.2 Semantic Knowledge Augmentation

3.2.1 Data Augmentation

To enable the model to detect unseen relations without labeled training instances, we first do data augmentation by translating a sentence from its original seen relation to a new unseen relation using an analogy. In the word level, we adopts 3CosMul \(^1\) (Levy and Goldberg, 2014) to get the candidates of new words \( w_u \):

\[
 w_u = \arg\max_{x \in V} \frac{\cos(x, r_u) \cdot \cos(x, w_s)}{\cos(x, r_s)} + \epsilon
\]

where \( V \) is the vocabulary set, \( \cos(\cdot) \) is the cosine similarity, \( r_u \) is the unseen relation name, \( r_s \) is the seen relation name, \( w_s \) is the word in seen class and \( \epsilon \) is a small number to prevent division by zero.

In the sentence level, we follow Algorithm 1 to translate a sentence of relation \( r_s \) into a new sentence of relation \( r_u \). To be more specific, we translate all nouns, verbs, adjectives, and adverbs in the seen sentence to a new sentence. We do the translation when the super-class of \( r_s \) and the super-class of two corresponding name entities in \( r_s \) are the same with the super-class of \( r_u \) and the super-class of two related name entities in \( r_u \). If the number of \( r_s \) that conforms to the above rules is larger than one, we take all the translated sentences and randomly select the same number as other seen relations to make a balanced training set.

\(^1\)We use top 10 similar words to return.
Algorithm 1: Sentence Generation for Unseen Relations

**Input**: sentence $x_i = [w_{i1}, \ldots, w_{it}]$, two name entities $e_{i1}$ and $e_{i2}$, original relation label sets $R_s$, target unseen relation label $r_u$

**Output**: sentence $x_i^u$ with relation $r_u$

for $r_u \in R_s$ do

if $S(r_u) == S(r_s)$ and $S(e_u) == S(e_s)$ then

for $w \in x_i$ do

if is_valid_pos(w) then

$w_u = 3CosMul(w, r_u, r_s)$

$x_i^u$.append($w_u$)

else

$x_i^u$.append(w)

else

Continue

end if

end for

end if

end for

return $x_i^u$

Algorithm 2: Virtual Label Generation

**Input**: word $w_i$, relation $r_c$, threshold $\tau$, number of hop $K$, Knowledge Graph $G$, number of virtual label $n$

**Output**: virtual label $r_v$

for $w_i \in V$ do

if $\frac{|w_i|_{r_c}}{\tau} \geq \tau$ then

$v_i = 0$, $v_2$, $v_3$, $v_{ave} = []$

if $w_i \in G$ then

$v_i = 1$

else

$v_i = 0$

end if

end for

for $k \in K$ do

hops = find_neighbors($w_i$) $\subseteq G$

if hops then

$v_2$.append(any(hops))

$v_3$.append(sum(hops))

$v_{ave}$.append(mean(hops))

else

$v_2$, $v_3$, $v_{ave}$.append(0)

end if

$\alpha_{w_i} = \sum_r v_r D(w_i)$

else

Continue

end if

end for

$\gamma_{w} = \frac{\alpha_{w} \cdot E(w_i) + \cdots + \alpha_{wn} \cdot E(w_n)}{\sum \alpha}$

return $r_v$

3.2.2 Prompts from Knowledge Graph

For relation classification, the core issue is to extract the relations related to the two given name entities from all aspects and granularities. For zero-shot tasks, we design prompts used as training instances to help train the model because there is no real training data available. From this perspective, we construct prompts based on external knowledge graph ConceptNet (Speer and Havasi, 2013), a knowledge graph that connects words and phrases of natural language with labeled edges, for zero-shot relation classification. Nodes in ConceptNet are entities, and edges connecting two nodes are semantic relations between the entities.

Because of the relation classification task, we wrap the input sequence with a template, which is a piece of natural language text. To be more specific, we build prompts as ‘$S(e_{i1})$ is [MASK] of $S(e_{i2})$’ \(^2\). The [MASK] here is a virtual label word $r_v$ representing the relation between $e_{i1}$ and $e_{i2}$.

Unlike using real words, we build the virtual label word that can primarily represent the relation in each sentence. Instead of building a virtual label word by simply using the mean vector of the top-K high-frequency words (Ma et al., 2021), we build our virtual label word based on a knowledge graph using the following strategy.

We firstly represent a relation $r$ as five sets of nodes in ConceptNet by processing the class label $r_c$, class hierarchy $S(r_c)$, class description $D(r_c)$ and hierarchy of two name entities $S(e_{i1})$ and $S(e_{i2})$. We consider whether a word $w_i$ is related to the members of the five sets above within $K$ hops or not. The value of $K$ is determined through grid search on the validation set. For each of the five sets above, we consider $v_1$ (whether $w_i$ is a node in $G$ in that set), $v_2$ (whether $w_i$’s neighbor is a node in $G$), $v_3$ (number of neighbors of $w_i$ in $G$). The above values associated with each set demonstrate the semantic distance of $w_i$ and the corresponding set. Detailed construction of virtual label $r_v$ is shown in Algorithm 2.

3.3 Model Architecture and Training

3.3.1 Instance Encoder

Figure 2 shows the architecture of the encoder used in this paper. We first tokenize and lemmatize all words in a sentence. There is a similar performance, so we put the prompts after each input sentence.
Figure 2: BERT-CNN Instance Encoder.

Our approach to generating the hidden state vectors of prompts is as follows.

We obtain the hidden state vectors of prompts where \(E\) is the embedding function, \(\text{ReLU}(\cdot)\) denotes the activation function, and \(\text{max}(\cdot, \cdot)\) denotes the max-pooling layer.

where \(c_i\) represents the prototype for each relation, \(f_\phi\) is the BERT-CNN encoder, \(h_i\) is the representation for each original or augmented sentence and \(p_i\) is the prompt embedding introduced in Sec. 3.2.2. The probabilities of the relations in \(R_s\) and \(R_u\) for a query instance \(x\) is calculated as:

where \(d(\cdot, \cdot)\) is Euclidean distance for two vectors.

4 Experiments

We conduct several experiments with ablation studies on three public datasets: FewRel (Han et al., 2018), Wiki-ZSL (Sorokin and Gurevych, 2017; Chen and Li, 2021) and NYT (Riedel et al., 2010) to show that our proposed model outperforms other existing state-of-the-art models, and our proposed model is more robust compared with the other models in zero-shot learning tasks.

4.1 Evaluation Settings

4.1.1 Dataset

In our experiments, we evaluate our model over three widely used datasets: FewRel (Han et al., 2018), Wiki-ZSL (Chen and Li, 2021) and NYT (Riedel et al., 2010). FewRel and Wiki-ZSL are two balanced datasets and NYT is an unbalanced dataset. The statistics of FewRel, Wiki-ZSL, and NYT datasets are shown in Table 1. We provide more detailed description in the Appendix.

Table 1: The statistics of each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#instances</th>
<th>#relations</th>
<th>avg. len.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FewRel</td>
<td>56,000</td>
<td>80</td>
<td>24.95</td>
</tr>
<tr>
<td>Wiki-ZSL</td>
<td>94,383</td>
<td>113</td>
<td>24.85</td>
</tr>
<tr>
<td>NYT</td>
<td>134,152</td>
<td>53</td>
<td>38.81</td>
</tr>
</tbody>
</table>

Each prototype is the average instance embeddings belonging to one relation:

where \(c_i\) represents the prototype for each relation, \(f_\phi\) is the BERT-CNN encoder, \(h_i\) is the representation for each original or augmented sentence and \(p_i\) is denotes the prompt embeddings introduced in Sec. 3.2.2. The probabilities of the relations in \(R_s\) and \(R_u\) for a query instance \(x\) is calculated as:

where \(d(\cdot, \cdot)\) is Euclidean distance for two vectors.
relations have not appeared in training data. For hyperparameter and configuration of ZS-SKA, we implement ZS-SKA with PyTorch and optimize it with SGD optimizer. The initial learning rate is selected via grid search within the range of \( \{le - 1, le - 2, le - 3, le - 4\}\) for minimizing the loss, the cosine similarity threshold is selected from 0 to 1 with step size 0.1. Table 5 in Appendix shows other parameters used in the experiment.

### 4.1.3 Baselines and Evaluation Metrics

We compare our proposed model to several state-of-the-arts models in zero-shot learning tasks. For clean FewRel and Wiki-ZSL datasets, we compare our model with CNN (Zeng et al., 2014), Bi-LSTM (Zhang et al., 2015), Attentional Bi-LSTM (Zhou et al., 2016), R-BERT (Wu and He, 2019), ESIM (Chen et al., 2017), CIM (Rocktäschel et al., 2016) and ZS-BERT (Chen and Li, 2021). The seven baselines above are reported by (Chen and Li, 2021). We also compare the robustness of our model with the most state-of-the-art re-implemented ZS-BERT. For noisy NYT dataset, we compare our model with the re-implemented CDN (Zeng et al., 2014), REDN (Li and Tian, 2020) and ZSLRC (Gong and Eldardiry, 2021). The evaluation metrics adopted in this paper are the Precision, Recall, and F1-score, similar to those used for the above baselines.

### 4.2 Results and Discussion

#### 4.2.1 Main Results

**Results on Balanced Datasets** The evaluation results of zero-shot learning on Wiki-ZSL and FewRel are shown in Table 2. We compare our proposed model ZS-SKA with models reported by (Chen and Li, 2021). Obviously, ZS-SKA significantly outperforms other state-of-the-art models on both balanced datasets. Our proposed ZS-SKA outperforms a recently proposed method (ZS-BERT) by 6.9% precision, 5.7% recall, and 3.9% F1-score on Wiki-ZSL, 9.8% precision, 13.5% recall, and 10.2% F1-score on FewRel. The performance improvement indicates that semantic knowledge augmentation is competitively more beneficial for relation classification than only incorporating text description of relations. To compare the robustness of ZS-SKA with the strongest baseline ZS-BERT, we conduct further experiments with different percentages (varying \( m \)) of unseen relations in Sec. 4.2.2.

**Results on Unbalanced Dataset** The experiment results on unbalanced dataset NYT by varying \( m \) unseen relations are shown in Table 3. To make fair comparisons, we use the same splitted NYT dataset and follow the same threshold schema provided by (Gong and Eldardiry, 2021). We remove data augmentation module and only implement the prompts generated through the knowledge graph as similar side information in ZSLRC model. Apparently, the proposed ZS-SKA achieves a substantial gain in precision, recall and F1-score over other baselines on the NYT dataset. When the number of unseen relations in the testing set becomes larger, the superiority of ZS-SKA gets more significant and robust. Such results indicate the effectiveness of leveraging prompts using virtual labels constructed from the knowledge graph instead of using keywords learned from the distribution of training data on the noisy dataset.

![Figure 3: F1 scores of models with different proportions of unseen relations.](image-url)
Table 2: Results with $m = 15$ on Wiki-ZSL and FewRel.

<table>
<thead>
<tr>
<th>Method</th>
<th>Wiki-ZSL Precision</th>
<th>Wiki-ZSL Recall</th>
<th>Wiki-ZSL F1</th>
<th>FewRel Precision</th>
<th>FewRel Recall</th>
<th>FewRel F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN (Zeng et al., 2014)</td>
<td>14.58</td>
<td>17.68</td>
<td>15.92</td>
<td>14.17</td>
<td>20.26</td>
<td>16.67</td>
</tr>
<tr>
<td>Bi-LSTM (Zhang et al., 2015)</td>
<td>16.25</td>
<td>18.94</td>
<td>17.49</td>
<td>16.83</td>
<td>27.62</td>
<td>20.92</td>
</tr>
<tr>
<td>Att Bi-LSTM (Zhou et al., 2016)</td>
<td>16.93</td>
<td>18.54</td>
<td>17.70</td>
<td>16.48</td>
<td>26.36</td>
<td>20.28</td>
</tr>
<tr>
<td>R-BERT (Wu and He, 2019)</td>
<td>17.31</td>
<td>18.82</td>
<td>18.03</td>
<td>16.95</td>
<td>19.37</td>
<td>18.08</td>
</tr>
<tr>
<td>ESIM (Chen et al., 2017)</td>
<td>27.31</td>
<td>29.62</td>
<td>28.42</td>
<td>29.15</td>
<td>31.59</td>
<td>30.32</td>
</tr>
<tr>
<td>CIM (Rocktäschel et al., 2016)</td>
<td>29.17</td>
<td>30.58</td>
<td>29.86</td>
<td>31.83</td>
<td>33.06</td>
<td>32.43</td>
</tr>
<tr>
<td>ZS-BERT (Chen and Li, 2021)</td>
<td>34.12</td>
<td>34.38</td>
<td>34.25</td>
<td>35.54</td>
<td>38.19</td>
<td>36.82</td>
</tr>
<tr>
<td><strong>ZS-SKA</strong></td>
<td><strong>41.03</strong></td>
<td><strong>40.12</strong></td>
<td><strong>38.13</strong></td>
<td><strong>45.34</strong></td>
<td><strong>51.67</strong></td>
<td><strong>47.02</strong></td>
</tr>
</tbody>
</table>

Table 3: Results with different $m$ values on NYT.

<table>
<thead>
<tr>
<th>Method</th>
<th>m=15</th>
<th></th>
<th></th>
<th>m=30</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>CDNN (Zeng et al., 2014)</td>
<td>27.94</td>
<td>44.10</td>
<td>33.72</td>
<td>10.17</td>
<td>25.62</td>
<td>14.23</td>
</tr>
<tr>
<td>REDN (Li and Tian, 2020)</td>
<td>66.52</td>
<td>65.47</td>
<td>66.98</td>
<td>57.19</td>
<td>56.80</td>
<td>56.99</td>
</tr>
<tr>
<td>ZSLRC (Gong and Eldardiry, 2021)</td>
<td>96.06</td>
<td>93.84</td>
<td>93.59</td>
<td>94.81</td>
<td>90.46</td>
<td>89.76</td>
</tr>
<tr>
<td><strong>ZS-SKA</strong></td>
<td><strong>96.23</strong></td>
<td><strong>94.68</strong></td>
<td><strong>94.42</strong></td>
<td><strong>95.91</strong></td>
<td><strong>90.38</strong></td>
<td><strong>91.27</strong></td>
</tr>
</tbody>
</table>

significantly than ZS-SKA. It is probably because prompts constructed by virtual labels contain the semantic information of unseen relations, which shorten the distance between the query sentence of an unseen relation with the prototype of such unseen relation.

Figure 4: Examples of denoising in virtual label construction on FewRel and Wiki-ZSL datasets.

**4.2.3 Case Study**

**Data Augmentation** Table 4 shows an example of the augmented data following the the translating rule in Sec. 3.2.1 on Wiki-ZSL dataset. Relation ‘place_of_birth’ is a seen class, and the four relations ‘place_of_death’, ‘residence’, ‘country’ and ‘educated_at’ are from unseen classes. We follow the data augmentation method introduced in Sec. 3.2.1 to generate augmented training instances for these unseen relations. We observe that if the super-class of the relation and the super-class of the two name entities are the same, i.e. ‘place_of_death’ with ‘place_of_birth’, ‘residence’ with ‘place_of_birth’, the generated sentences have a good quality with the name entities having the unseen relations. If the super-class of the relation or super-class of the two name entities of unseen relation is different from that of the seen relation, i.e. ‘country’ with ‘place_of_birth’, ‘educated_at’ with ‘place_of_birth’, though the generated sentences contain the tone of the target (unseen) relation, such as the words in blue, the original two name entities do not have the target unseen relation. For example, the generated sentence of relation ‘country’ in Table 4 can be explained that Arsenal is from a European country, but such relation is lost between the original two name entities ‘Rich’ and ‘Arsenal’. Therefore, we follow the rule of using the relation and name entities from the same super-class with that of unseen relations to generate high-quality augmented instances for training in zero-shot settings.

**Virtual Label Construction** Figure 4 shows an example of the ranking top ten components of the constructed virtual label before denoising and after denoising. The virtual labels shown in Figure 4 are generated by Algorithm 2 proposed in Sec. 3.2.2.
Table 4: Examples of sentence generation from seen relations by data augmentation. Words in red are name entities for each sentence. $S(\cdot)$ denotes the super-class of the relation or name entities.

<table>
<thead>
<tr>
<th>Relation</th>
<th>$S(r)$</th>
<th>$S(e_1)$</th>
<th>$S(e_2)$</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>place_of_birth</td>
<td>location</td>
<td>person</td>
<td>location</td>
<td>Jessica (born in Manchester) is a British track and field athlete who competes in the heptathlon.</td>
</tr>
<tr>
<td>place_of_death</td>
<td>location</td>
<td>person</td>
<td>location</td>
<td>Johnson (died in Liverpool) is a Military track and field athlete who competed in the decathlon.</td>
</tr>
<tr>
<td>residence</td>
<td>location</td>
<td>person</td>
<td>location</td>
<td>Mansion (resided in Villa) is a Colonial residence and peri alumnus who dominates in the decathlon.</td>
</tr>
<tr>
<td>country</td>
<td>location</td>
<td>location</td>
<td>location</td>
<td>Rich (retired in Arsenal) is a European track and field athlete who competes in the decathlon.</td>
</tr>
<tr>
<td>educated_at</td>
<td>act</td>
<td>person</td>
<td>org.</td>
<td>Jess (motivate in Liverpool) is a British aims and professional athlete who educated in the decathlon.</td>
</tr>
</tbody>
</table>

The red words in Figure 4 (a) are irrelevant to the relation ‘religion_of’. After we refine the virtual label sets using the distance metric, these irrelevant words are filtered out in our virtual label sets, removing the noise in the knowledge graph.

4.2.4 Hyperparameter Sensitivity

We examine how some primary hyperparameters, including threshold $\tau$ for denoising virtual label sets and the number of virtual labels $n$ in Algorithm 2 affect the performance of ZS-SKA. By fixing $m = 15$ and varying $\tau$ and $n$, the results in terms of F1 scores and Accuracy on NYT, FewRel and Wiki-ZSL datasets are exhibited in Figure 5. We find that parameters $\tau$ and $n$ affect the noisy dataset more than the clean and balanced dataset. We conjecture that because both $\tau$ and $n$ are used for removing noise and getting more related semantic information in prompts construction, the noise in prompts may impact more on noisy datasets because noisy datasets are more sensitive to the noise.

If the threshold $\tau$ is between 0.5 and 0.6, it achieves the best performance on all three public datasets. This is reasonable that when $\tau$ is too low, most connected nodes in the knowledge graph are used to construct virtual label words. Thus, when building the prompts for each relation, it is more likely to bring the noise to the relation class. In contrast, when $\tau$ gets too high, some highly related nodes are filtered out to construct virtual labels. We would suggest setting $\tau$ between 0.5 to 0.6 to derive satisfying results across datasets. As for the number of words $n$ to construct virtual labels, we find that increasing the number of related words $n$ to construct virtual labels can achieve better performance. It is reasonable because, including more nodes (words) from the knowledge graph to construct the virtual label representing the relation information, more semantic knowledge information is contained, leading to a shorter distance between the query sentence embedding with the prototype constructed from the prompts.

Figure 5: Effects on varying threshold $\tau$ and number of virtual labels $n$ on NYT, FewRel and Wiki-ZSL datasets.

5 Conclusion and Future Work

In this paper, we propose a prompt-based ZS-SKA utilizing semantic knowledge augmentation to detect unseen relations with no corresponding labeled data available for training to tackle with zero-shot relation classification task. The experiments show that with augmented instances and prompts generated through a knowledge graph, ZS-SKA outperforms other state-of-the-art models under zero-shot learning. We have also conducted extensive experiments to study different aspects of ZS-SKA, from ablation study to hyperparameter sensitivity, and demonstrate the effectiveness and robustness of our proposed model. We plan to explore the following directions in future work: (1) Different ways of instance generation and prompt designs for semantic augmented data. (2) Better approaches for constructing virtual labels in the prompt template.
References


A Dataset Description

In the following, we describe each dataset in detail.

- **FewRel (Han et al., 2018).** The FewRel dataset is a human-annotated balanced few-shot RC dataset consisting of 80 types of relations, each of which has 700 instances.

- **Wiki-ZSL (Chen and Li, 2021).** The Wiki-ZSL dataset is a subset of Wiki-KB (Sorokin and Gurevych, 2017), which filters out both the 'none' relation and relations that appear fewer than 300 times.

- **NYT (Riedel et al., 2010).** The NYT dataset was generated by aligning Freebase relations with the New York Times corpus (NYT). There are 53 possible relations in total. It is an unbalanced noisy dataset because all the relations have a different number of sentences.

B Parameter Settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Embedding Dimension</td>
<td>768</td>
</tr>
<tr>
<td>Hidden Layer Dimension</td>
<td>300</td>
</tr>
<tr>
<td>Sentence Max Length</td>
<td>128</td>
</tr>
<tr>
<td>Convolutional Window Size</td>
<td>3</td>
</tr>
<tr>
<td>Batch Size</td>
<td>4</td>
</tr>
<tr>
<td>Initial Learning Rate $\alpha$</td>
<td>0.01</td>
</tr>
<tr>
<td>Number of Hops $K$</td>
<td>1</td>
</tr>
<tr>
<td>Similarity Threshold $\tau$</td>
<td>0.6</td>
</tr>
<tr>
<td>Number of Virtual Label $n$</td>
<td>5</td>
</tr>
</tbody>
</table>