PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction

Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, Laurent-Walter Goix

PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction

Pierre-Yves Genest[1, 2], Pierre-Edouard Portier[2], Előd Egyed-Zsigmond[2], and Laurent-Walter Goix[1]

[1]Alteca, 88 Boulevard des Belges, 69006 Lyon, France
[2]Univ Lyon, INSA Lyon, LIRIS, CNRS UMR5205, 20 Avenue Einstein, 69621 Villeurbanne, France
{*pygenest,lwgoix*}*@alteca.fr*
{*pierre-edouard.portier,elod.egyed-zsigmond*}*@insa-lyon.fr*

October 17, 2022

Abstract

Unsupervised Relation Extraction (RE) aims to identify relations between entities in text, without having access to labeled data during training. This setting is particularly relevant for domain specific RE where no annotated dataset is available and for open-domain RE where the types of relations are *a priori* unknown. Although recent approaches achieve promising results, they heavily depend on hyperparameters whose tuning would most often require labeled data. To mitigate the reliance on hyperparameters, we propose PromptORE, a "Prompt-based Open Relation Extraction" model. We adapt the novel prompt-tuning paradigm to work in an unsupervised setting, and use it to embed sentences expressing a relation. We then cluster these embeddings to discover candidate relations, and we experiment different strategies to automatically estimate an adequate number of clusters. To the best of our knowledge, PromptORE is the first unsupervised RE model that does not need hyperparameter tuning. Results on three general and specific domain datasets show that PromptORE consistently outperforms state-of-the-art models with a relative gain of more than 40% in B^3, V-measure and ARI. Qualitative analysis also indicates PromptORE's ability to identify semantically coherent clusters that are very close to true relations.

Keywords— unsupervised relation extraction, open relation extraction, natural language processing, prompt-tuning

Disclaimer

This paper was published as a part of the Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.

This document is the accepted/author version, **not** the published version. The definitive version is accessible at `https://doi.org/10.1145/3511808.3557422`.

1 Introduction

Information Extraction models aim to extract the meaningful information from text, that is, entities and relations between these entities. The resulting network of relations can then be transformed into knowledge graphs that are used in multiple downstream tasks such as recommender systems [16], logical reasoning [6] or question answering [22]. Information Extraction is usually seen as a two-step process: 1. Named Entity Recognition and 2. Relation Extraction. In this paper, we focus on Relation Extraction, which consists in identifying the relation between two entities in the context of a piece of text.

Relation Extraction (RE) is often seen as a supervised task [24], thus relying on datasets labeled with a predefined set of relations. However, this setting can be restrictive for some applications, especially domain-specific RE lacking annotated data or open-domain RE where we do not know in advance the relations expressed in the dataset. Therefore, more flexible paradigms have been proposed such as distant-supervision [38, 43], which tries to automatically annotate data; few-shot learning [40], which learns from a very small set of labeled instances; or unsupervised learning. In particular, unsupervised RE (also called OpenRE) does not require a training dataset with labeled relations and assume no prior knowledge about expected types of relations.

Several recent OpenRE approaches obtain interesting results on datasets containing tens or hundreds of relation types [21, 62, 33]. They often try to compute a vector representation of the relation expressed in the sentence (also called *relation embedding*) and then cluster all the embeddings to identify groups of similar relations. Most of these methods rely on hyperparameters (e.g. number of epochs, regularization, early stopping, number of relations types, ...) that have a significant impact on their overall performance. However, tuning these hyperparameters most often requires access to labeled data, thus limiting the applicability of such models in a real-world unsupervised scenario.

We therefore propose **PromptORE**, a "Prompt-based Open Relation Extraction" model, which relies on one hyperparameter at most: the target number k of relation types to be extracted. Our experiments show that even when no educated guess can be made about k, an efficient estimate can easily be obtained in an automatic way. Thus, to the best of our knowledge, **PromptORE** is the first proposal for an unsupervised Relation Extraction system that can operate in a fully unsupervised setting.

To achieve this, we first compute, for each instance (a piece of text) of a dataset, a relation embedding that represents the relation expressed in the instance. Contrary to previous approaches that fine-tuned BERT [21, 62, 71], we use the novel prompt-tuning paradigm. Prompt-tuning replaces the usual training by designing a prompt (i.e., a text that is inputted to BERT), able to elicit as much information as possible from the Pretrained Language Model. Prompt-tuning is already used in few-shot RE [35, 5, 15, 56]. We propose to go further and adapt this paradigm to work in a fully unsupervised way. Prompt-tuning has many benefits: 1. it does not involve training or fine-tuning BERT, thus removing a significant number of hyperparameters, 2. the proposed encoder is extremely simple, yet 3. we show that these prompt-based relation embeddings provide better results than current state-of-the-art methods. Usual clustering algorithms are then applied to group together the embeddings in order to discover relation types.

Let us summarize our main contributions:

- We propose **PromptORE**, a novel OpenRE model that minimizes the number of hyperparameters and provides clear ways to tune its only hyperparameter k in a strict unsupervised setting.
- We adapt the prompt-tuning paradigm to an unsupervised setting, which allows

us to leverage more expressive embeddings than previous entity-pair representations [21, 62, 34].

- We show that this model outperforms consistently previous state-of-the-art approaches on three different datasets, covering both general and specific domains. We also demonstrate that the predicted clusters are semantically coherent and very close to the true relations.

2 Related Work

For the sake of clarity, we use *relation* and *relation type* interchangeably, and we define them as a concept linking two entities, for example `married_to`, `born_in`, `located_in`; but we distinguish *relation instance*, which represents the realization of a relation, that is, a piece of text expressing the relation.

Relation Extraction aims to discover the binary relation that links two entities mentioned in a text. RE allows to extract triples of the form (**e1**, relation, **e2**) that can be used thereafter to build for instance a knowledge graph. Even though Relation Extraction from documents is the most general paradigm, the majority of models focus on extraction within a single sentence ignoring inter-sentence relations [18, 17]. Recent approaches follow a two-step process [17, 28, 34, 53]:

1. Relation Embedding, which computes a vector representation of the relation instance,
2. Relation Classification.

To compute relation embeddings, word-embedding models are often used, such as GLoVe [39], ELMO [41], Bi-LSTM embeddings [34], or more recently BERT embeddings [59, 9]. In the general case, relation classification is seen as a supervised task therefore needing labeled datasets, where entities have been extracted and labels describing the relation are available.

Supervised Relation Extraction Recent models implement a joint entity and relation extraction scheme [24]. Luan et al. [34] propose a multi-task learning approach that simultaneously optimizes a BERT-based entity extractor, relation extractor and coreference resolution model. Wadden et al. [60] add event triggers and roles detection into the same multi-task framework. Lin et al. [28] improve these models by incorporating external constraints on entity types and relations. Zhong et al. [73] show nevertheless that a simple pipelined approach outperforms complex multi-task models, thanks to a markup-based encoding of the sentence. However, these supervised approaches rely on large labeled datasets, available only for generic language corpora and a few specialized domains. Therefore, other works aim at reducing this reliance on such annotated data.

Distantly-supervised Relation Extraction Distant-supervision [38, 43] tackles this problem by automatically annotating texts based on external knowledge bases. This annotation process creates large scale datasets that are characterized by a high level of noise. Zhang et al. [69] identify two types of noise: intra-dictionary bias (spurious entity/relation annotations due to wrong entity linking) and inter-dictionary bias (non exhaustiveness of knowledge bases, meaning that some entities/relations cannot be labeled). Most works focus on trying to reduce and mitigate these biases [68, 70, 69, 72]. However, distant-supervision does not apply to very specific domains that lack large knowledge bases.

Few-Shot Relation Extraction This approach aims to learn a relation extraction model from the least amount of labeled data. Models focus on relation classification: they suppose the existence of a relation between two entities [40], ignoring the case of sentences mentioning unrelated entities. Snell et al. [52] propose to use prototypical networks to determine a prototype for each relation, and predict the relation by measuring the distance between the relation instance embedding and each prototype. Ren et al. [42] and Zhao et al. [71] propose to improve this method using transfer-learning with general domain labeled datasets.

Very recent few-shot methods consider the use of prompt-tuning with BERT and more broadly Pretrained Language Models (PLMs) [13], as it allows for more efficient learning in low-resource setting [30]. Prompt-tuning replaces fine-tuning by designing a "prompt", that is, a piece of text containing the special `[MASK]` token, and ask a PLM such as BERT to predict the embedding of this `[MASK]` token. This embedding is then compared with a set of target tokens (that can be seen as relation prototypes) to determine the relation expressed in this sentence [35, 5, 15, 56]. Efforts are focused in optimizing the prompt $\mathcal{P}$ and selecting a set of target tokens effective at representing the relations. In particular, Jiang et al. [25] propose text-mining and paraphrasing-based methods to generate prompts.

The current major limitation with few-shot RE is the closed-world hypothesis, which stipulates that relations must be known in advance (although recent papers start to explore *none-of-the-above* prediction [14, 35]).

Unsupervised Relation Extraction Unsupervised RE (or OpenRE [3]) aims to extract relations without having access to a labeled dataset during training. Methods can be divided in two subgroups: (1) triples extraction and (2) relation typing. Banko et al. [3] extract triple candidates using syntactic rules and refine the candidates with a trained scorer. Saha et al. [48] propose to simplify conjunctive sentences to improve triples extraction. More recently, neural networks and word-embedding were applied to solve this task [54, 8], requiring a general domain annotated dataset to pretrain their model. Finally, Roy et al. [47] propose an ensemble method to aggregate results of multiple OpenRE models. These triples extraction approaches rely on surface forms, which makes it hard for them to group instances that express the same relation using very different words and syntax.

To solve this problem, Yao et al. [63] propose instead to learn a relation classifier, using Latent-Dirichlet Allocation [4], a generative probabilistic model. The majority of these relation typing methods rely on relation embeddings: first, they compute an embedding, which encodes the underlying relation, second they use this embedding to identify groups of relation instances. The earliest methods use syntactic and semantic features [63, 64, 37]. Elsahar et al. [10] add word-embedding features based on GloVe [39], apply dimensionality reduction methods, and an agglomerative clustering model to identify clusters of relation instances. Marcheggiani et al. [37] use a fill-in-the-blank task: they mask one entity and try to predict it using a Variational Auto-Encoder (VAE) [26], proving the benefit of generating a supervision signal. This method is further improved by adding two regularization losses to limit overfitting [51] and by finding a more effective formulation of the VAE task [65]. Tran et al. [58] however succeed to outperform VAE approaches only using entity types as their relational embeddings.

Hu et al. [21] adopt an other supervision signal: they compute pseudo labels using a k-means clustering on relation embeddings, and train a classifier to reproduce these pseudo labels, allowing them to fine-tune a BERT model. As an alternative, Wu et al. [62] propose to learn a distance metric representative of the relations (using Siamese neural networks [7]), to compare pairs of instances. This metric is learned on an annotated dataset, and applied to unlabeled data to identify instances expressing similar relations. Lou et al. [33] use ranked list loss [61] as an alternative to Siamese

neural networks. Finally, a tendency of recent unsupervised RE methods is to use transfer-learning: learning some relation embeddings or metrics on general domain annotated datasets and try to adapt them to unsupervised data [71, 33, 62]. Compared to triples extraction, relation typing assumes that there is always a relation between the two entities, which can be seen as a limitation.

To allow evaluation of such OpenRE models, previous works tend to train them on labeled datasets and compare their predictions with ground truth relations using external clustering evaluation metrics such as V-measure [45], Adjusted Rand Index [23, 55] or B^3 [2].

Are current OpenRE models truly unsupervised ? Although Open-RE models extract relations from unannotated datasets, we argue that they are not truly unsupervised approaches: the main problem is hyperparameter tuning. All these approaches rely extensively on hyperparameters that need to be adjusted: number of epochs/iterations [37, 65, 58, 21, 62, 33], learning rate, regularization [51], entity types [58], early-stopping [58], etc., and most importantly the number of relations k the model is supposed to extract [37, 65, 58, 21, 62, 33, 51, 10]. In a real unsupervised setting these hyperparameters are extremely hard to determine, and cited papers do not present satisfactory methods to estimate them without labeled data. Therefore, we conclude that these mentioned approaches are not fully unsupervised when it comes to hyperparameter tuning, which in our opinion, restricts their use in a real-world application.

As a result, it motivates us to define more precisely the unsupervised RE setting as *learning a RE model and tuning its hyperparameters using only unlabeled data.*

3 Proposed Model

PromptORE aims to extract the binary relation r between two already known entities **e1** and **e2** present in the same sentence[1]. More precisely, as we follow an unsupervised setting, the first objective of **PromptORE** is to group instances expressing the same relation r, without having access to labeled data during training and hyperparameter tuning. Our second objective is to minimize the number of hyperparameters needed by **PromptORE** and to provide clear procedures to adjust them without annotated data.

To achieve these goals, we suppose we have access to a dataset $\mathcal{D}$ (see Figure 1) containing instances with the following properties:

- An instance is described with a triple $(\mathbf{S}, \mathbf{e1}, \mathbf{e2})$, where $\mathbf{S} = [t_0, ..., t_{s-1}]$ is the instance text composed of tokens[2], $\mathbf{e1} = [t_{start(\mathbf{e1})}, ..., t_{end(\mathbf{e1})}]$ and $\mathbf{e2} = [t_{start(\mathbf{e2})}, ..., t_{end(\mathbf{e2})}]$ are two entities identified by their indexes in $\mathbf{S}$.
- We suppose that **e1** and **e2** have already been extracted (but not typed).
- In the instance text $\mathbf{S}$, **e1** and **e2** are linked by a binary relation r. As previous approaches, we do not consider the case where there is no relation between **e1** and **e2**.
- We do not have access to any relation label during training and hyperparameter tuning.

$\mathcal{R}$ is the set of the k relations contained in $\mathcal{D}$. We consider that we have no information about the relations in $\mathcal{R}$ (e.g. their labels, their linked entity types, etc.). Regarding k, it can either be given by the user or automatically estimated by methods described in section 3.2.

[1] As previous works, we focus on sentence RE, even though we are aware that some relations may be missed.

[2] *Token* as defined by BERT [9]: punctuation, word or part of word.

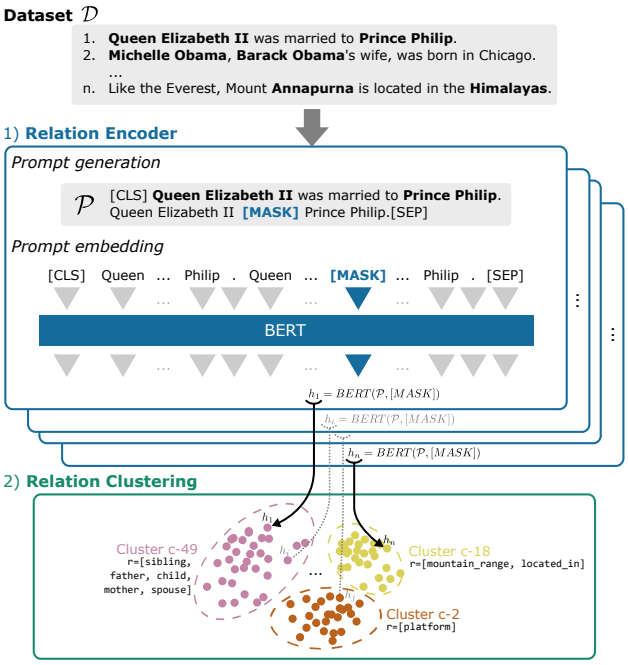

Figure 1: Overview of PromptORE.

As shown in Figure 1, **PromptORE** is composed of two main modules (similarly to [21, 71]):

1. *Relation Encoder*. This module computes a vector representation of the relation that is expressed in the current instance. To do that, we apply a modified prompt-tuning method to leverage BERT embeddings.
2. *Relation Clustering*. It clusters the relation embeddings of the whole dataset, in order to identify groups of instances that are expected to express the same type of relation[3].

3.1 Relation Encoder

This module aims to compute a vector representation (or *relation embedding*) of the relation expressed between **e1** and **e2** in the current sentence **S**. We want this relation embedding to be representative of the underlying relation: if the relation embeddings of two instances are close (relative to a certain distance metric), these instances convey, most probably, the same relation. In other words, the Relation Encoder aims to abstract the notion of relation instance, to provide embeddings that are easier to compare.

In recent papers [21, 71, 29, 33, 65, 73, 60], relation embeddings are computed using Pretrained Language Models such as BERT [9, 59] or RoBERTa [31]. BERT (and RoBERTa) takes as input some tokenized text, and computes for each input token an embedding, which is representative of the token itself and its context of use. BERT also contains a Masked Language-Model (MLM) head, which allows it to predict the most probable tokens associated with an embedding. In addition to "real word" tokens, BERT uses some special tokens:

[3]In practice, clusters may not be perfectly pure: they may contain instances expressing different relations, e.g. clusters c-18 or c-49 in Figure 1 or Table 4.

- `[CLS]` and `[SEP]`. By convention, `[CLS]` needs to be inserted at the start of the text, and `[SEP]` indicates the end of the text.
- `[MASK]`. It represents a token that is hidden/unknown, and BERT will try to compute a satisfactory embedding. Then, using the MLM head, BERT can predict the most probable tokens. Thanks to this `[MASK]` token, BERT can auto-complete sentences or generate text.

We define $h = \text{BERT}(\mathbf{S}, \texttt{[MASK]})$ as the BERT-embedding of the token `[MASK]`, in the context of $\mathbf{S}$ (a piece of text that must contain exactly one `[MASK]` token). Finally, for the sake of simplicity, variables/parameters (written with a bold name), are automatically replaced by their value in a quoted string. For example, if $\mathbf{a} = \textit{One}$, `"a, Two"` corresponds to *One, Two*.

Previous works [21, 71, 29, 33, 65, 73, 60] use an entity-pair representation paradigm to compute relation embeddings. We however opt for another technique: prompt-based encoding.

Prompt-Tuning The idea behind prompt-tuning is to benefit from BERT's ability to predict masked tokens (`[MASK]` tokens). It is already used in few-shot RE by [15, 5, 35, 56]. It can be summarized as follows:

1. Design a prompt $\mathcal{P}$, which is a sequence of tokens that includes one `[MASK]` token. For instance, Lv et al. [35] use the template $\mathcal{P}(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) =$ `"[CLS] S In this sentence, e1 is the [MASK] of e2.[SEP]"`.
2. Identify a set of label tokens $\mathcal{L}$ that represents each relation.
3. Predict the `[MASK]` embedding in the context of $\mathcal{P}$ using BERT: $h = \text{BERT}(\mathcal{P}, \texttt{[MASK]})$.
4. With this embedding, compute the probability to predict each label token $l \in \mathcal{L}$ thanks to the MLM head of BERT.
5. Select the relation represented by the label token l with the highest probability.

Prompt-tuning does not require to fine-tune BERT, but necessitate to design an optimal prompt $\mathcal{P}$ and a set of label tokens $\mathcal{L}$. They are usually adjusted using a labeled dataset and therefore cannot be applied directly to our unsupervised Relation Encoder.

Unsupervised Prompt-based Relation Encoder To adapt prompt-tuning for unsupervised RE, we propose to remove the set of label tokens $\mathcal{L}$, and use the simplest prompt $\mathcal{P}$ possible. Our proposed prompt template is:

$$\mathcal{P}(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) = \texttt{"[CLS] S e1 [MASK] e2.[SEP]"} \tag{1}$$

with $\mathbf{S}$ the instance text, $\mathbf{e1}$ (resp. $\mathbf{e2}$) the text of the first (resp. second) entity. For example, given the sentence $\mathbf{S} = \textit{Queen Elizabeth II was married to Prince Philip.}$, and entities $\mathbf{e1} = \textit{Queen Elizabeth II}$ and $\mathbf{e2} = \textit{Prince Philip}$, the prompt is:

$$\mathcal{P} = \textit{[CLS] Queen Elizabeth II was married to Prince Philip. Queen Elizabeth II [MASK] Prince Philip.[SEP]}$$

As we remove $\mathcal{L}$, we decide to use the `[MASK]` BERT embedding as our relation embedding. Thus, the Relation Encoder process is the following (as shown in Figure 1):

1. Apply the template defined in eq. (1) to generate a prompt for the current instance.
2. Predict the embedding of the `[MASK]` token with BERT, and use it as our relation embedding: $h = \text{BERT}(\mathcal{P}, \texttt{[MASK]})$.

Alternative Prompts If we analyze $\mathcal{P}$, we can see that BERT will likely fill the `[MASK]` token with a verb, as it is trained to produce grammatically correct sentences. It raises questions: can all relations be expressed with a *verb*, and with a *single* word? We did an analysis on the 9 892 Wikidata relations with surface forms:

- More than 75% of relations need 2 words or more to be expressed (e.g. acceptable surface forms for `birth_place` are *born in* or *the birth place of*).
- Surface forms usually contain a root word (noun, verb) accompanied by tool words. 92% of the root words are nouns, and only 6.7% verbs. The most common tool words are: *of, in, by, the, a, to.*

It can be therefore interesting to consider alternative prompts that encourage BERT to predict a noun ([35] also focus on noun prediction). In addition, Lv et al. [35] introduce a prefix to their prompt: `"In this sentence"`. Therefore, we define alternative prompt templates aiming at predicting noun, and with various prefixes:

$\mathcal{P}'_1(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) =$ `"[CLS] S e1 is the [MASK] of e2.[SEP]"`

$\mathcal{P}'_2(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) =$ `"[CLS] S In this sentence, e1 is the [MASK] of e2.[SEP]"`

$\mathcal{P}'_3(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) =$ `"[CLS] S We deduce that e1 is the [MASK] of e2.[SEP]"`

$\mathcal{P}'_2$ is the same as [35]. However, we cannot choose the optimal prompt from the previous ones nor use automatic methods to generate prompts (such as [25]) as they require access to labeled data. Therefore, the main results of **PromptORE** are computed using $\mathcal{P}$, the simplest prompt of all. In a second phase, we will analyze **PromptORE**'s performances with these alternative prompts.

3.2 Relation Clustering

We can measure the similarity between two BERT embeddings using an euclidian distance as these embeddings are normalized [9]. Similarly to previous works [10, 21, 71], we cluster the relation embeddings computed on the entire dataset $\mathcal{D}$ to find groups of instances, and we expect these clusters to be good candidate relations.

K-Means Clustering If we know in advance the number of relations k, we propose to use a simple k-means clustering [32, 36].

Clustering without k The most general case is however that we do not know k. To tackle this problem, we can take two points of view: use clustering models that do not require a predefined number of clusters, or estimate k automatically and use it with a regular clustering method.

For the first point of view, multiple models are available, the main ones being Agglomerative Clustering (HAC) [50], DBSCAN [11], OPTICS [1] or Affinity Propagation [12]. Nevertheless, most cannot be applied in our case: DBSCAN and HAC need other hyperpameters (such as density) and Affinity Propagation does not scale well to big datasets. Therefore, we propose to use OPTICS.

To estimate the number of clusters, we propose to implement the Elbow Rule [57]. We select the silhouette coefficient [46] as our internal metric[4] to measure the quality of the clustering. Intuitively, we expect that increasing the number of clusters will improve the value of the silhouette coefficient since there are more parameters to explain the data. However when we have more clusters than the actual number of relations, the silhouette will most likely grow more slowly as we can only subdivide actual relations. The Elbow Rule tries to find the "elbow", which is the optimal trade-off between a reasonable number of clusters and a high silhouette coefficient. It can be done visually, but automatic methods are also available [27]. With this estimation, we can use any clustering algorithm, in particular k-means.

[4] A metric that does not rely on external data such as labels.

4 Experiments

4.1 Datasets & Evaluation Metrics

To evaluate **PromptORE** we exploit labeled datasets, the labels being only used during evaluation. The first dataset we choose is FewRel [19][5]. This dataset is composed of text taken from Wikipedia pages that has been automatically annotated by aligning the text with Wikidata triples (distant-supervision setting), then manually checked for each instance. FewRel contains 80 relations, with 700 instances each (20 others relations are available in the test set, which is kept private). Therefore the dataset is composed of 56 000 instances. One important fact is that the dataset contains at most one instance for each pair of entities.

FewRel is a general domain dataset, but we also want to evaluate **PromptORE** on more specific domains. Our second dataset is FewRel NYT [14][5]. This time, the text is taken from newspapers articles of the New-York Times. It is also automatically annotated using Wikidata, and manually checked. FewRel NYT contains 25 different relations, with 100 instances each.

Our third dataset is FewRel PubMed [14][5]. The text comes from PubMed, a database of biomedical literature. It is also automatically annotated (this time with the UMLS knowledge base) and manually checked. It is composed of 10 relations, with 100 instances each.

Traditional classification metrics such as accuracy, precision, recall or f1-score cannot be used to evaluate and compare **PromptORE**'s performances as there is no direct link between our cluster ids and the true relation ids. Therefore, similarly to previous works [63, 51], we use three external clustering metrics: B^3 [2], V-measure (V) [45] and Adjusted Rand Index (ARI) [23, 55]. They each take a different point of view: ARI is based on pairwise similarity (enumerating all pair of instances), B^3 on one instance versus the dataset and the V-measure on clusters.

ARI is adjusted for chance, meaning that a random clustering will reliably lead to a score close to 0.

V-measure defines the notions of homogeneity and completeness of clusters. A cluster is homogeneous if it contains only instances of the same relation, and a cluster is complete if it contains every instances of a relation. The V-measure corresponds to the harmonic mean of homogeneity and completeness.

Finally, B^3 provides definitions for recall and precision, allowing to compute a f1-score. V-measure tends to penalize more small impurities in a pure cluster than impurities in a less pure cluster where B^3 has a more linear behavior [51].

4.2 Baselines

We compare **PromptORE** with the state-of-the-art (SOTA) approach **SelfORE** [21], and two previous approaches based on Variational Auto-Encoders (**Etype+** [58] and **UIE-PCNN** [51]).

SelfORE encodes instances with BERT, clusters these embeddings with an adaptive clustering method to generate pseudo labels that are finally used to train a classifier.

UIE-PCNN encodes instances with a Piecewise Convolutional Neural Network (PCNN) [66], and uses a Variational Auto-Encoder (VAE) to classify instances in an unsupervised way. Additionally they propose two regularization losses (skewness and dispersion) to fight against the VAE's tendency to predict a single relation or a uniform distribution. Since **UIE-PCNN** relies on PCNN, an older embedding method, we propose to replace it with a BERT model (similarly to [58]), and we call this method **UIE-BERT**.

[5]Data can be downloaded from: `https://github.com/thunlp/FewRel`.

Etype+ shows that using only entity types to encode instances provides better results than **UIE-PCNN**. They propose a simple typing schema for the entities: *Organization, Person, Location, Miscellaneous*. We expect the performances to be lower on domain-specific datasets (FewRel NYT and PubMed).

Finally, let us recall some metrics properties to interpret the experimental results: if a model always predicts the most frequent class V-measure and ARI will be equal to 0. If a model predicts a random distribution, ARI will be close to 0 (as ARI is adjusted for chance).

4.3 Implementation Details

All baselines are trained with the hyperparameter values determined by their authors. For **SelfORE**, we use their publicly available implementation, and for **Etype+** and **UIE-PCNN** we use the implementation of Tran et al. [58]. The baselines are trained knowing the correct number of relations k (i.e., 80 for FewRel, 25 for FewRel NYT and 10 for FewRel PubMed).

For **PromptORE**, we suppose we do not know k, except in section 5.1. Besides, we use the `bert-base-uncased` model to initialize BERT's weights. We also use a model with RoBERTa embeddings (using the `roberta-base` pretrained model). There are no hyperparameters to adjust.

We use the `scikit-learn` implementation of V-measure and Adjusted Rand Index, and the Hu et al. [21] implementation of B^3.

5 Results

5.1 Comparison with Previous SOTA Models

In this section only, to allow a fairer comparison with previous approaches, **PromptORE** knows k, the number of different relations, and a k-means clustering is used. Table 1 shows the results of the models on our three datasets. **PromptORE** consistently outperforms **SelfORE**, the previous state-of-the-art method, with approximately 19% more in B^3, 18% in V-measure and 19% in ARI on FewRel. It represents a relative gain in performance of more than 40%. The performance gap is even more important with **UIE-BERT**, **UIE-PCNN** and **EType+**. We observe similar conclusions with the two other datasets.

When we look more closely, we notice that **UIE-BERT** [51] obtains very poor results on all three datasets: it always predicts the same relation. Strangely, the authors of [51] proposed two regularization losses to avoid precisely this situation, but this problem with **UIE-PCNN/BERT** is also observed by [58, 65]. We believe this is due to the hyperparameter that controls the balance between classification and regularization losses, which needs to be fine-tuned specifically for each dataset.

SelfORE has been evaluated on the FewRel dataset multiple times [71, 67, 33], and it seems at first glance that the results obtained by these papers are better than ours (with an F1 B^3 between 45-55% instead of 25-30%). However, in these cases **SelfORE** was evaluated on a test set of FewRel with only 16 relations and 11 200 instances [33]; or a subset of 1 600 instances with 16 relations [71, 67]. Using the same sampling procedure, we were able to reproduce their results; but we do not use this setting in our evaluation, as it is a simpler task than FewRel with its 80 different relations.

Finally, we notice a very small difference in performance between BERT and RoBERTa embeddings with **PromptORE**. In practice both PLMs are well suited to provide precise results, and we decide to use BERT embeddings for the next parts of this paper.

Table 1: Results of PromptORE and previous SOTA models on three datasets. PromptORE knows the number of relations k.

Dataset		Model	B^3			V-measure			ARI
			Prec.	Rec.	F1	Hom.	Comp.	F1	
FewRel [19]	$k = 80$	UIE-PCNN [51]	5.20	6.78	5.89	21.1	21.6	21.3	4.86
		UIE-BERT	1.25	100	2.47	0	100	0	0
		EType+ [58]	7.46	7.99	13.7	33.3	79.1	47.9	8.44
		SelfORE [21]	24.4	36.3	29.2	50.4	56.6	53.2	24.4
		PromptORE (RoBERTa)	47.8	47.9	47.9	71.2	72.5	**71.8**	**43.7**
		PromptORE (BERT)	48.7	48.8	**48.8**	71.0	72.7	**71.8**	43.4
FewRel NYT [14]	$k = 25$	UIE-PCNN [51]	7.31	27.1	11.5	9.58	15.8	11.9	3.09
		UIE-BERT	4.00	100	7.77	0	100	0	0
		EType+ [58]	11.0	92.6	19.6	23.0	84.9	36.2	7.82
		SelfORE [21]	32.4	48.1	38.7	50.0	58.9	54.1	26.8
		PromptORE (RoBERTa)	62.6	65.3	63.9	75.7	78.1	76.8	**57.3**
		PromptORE (BERT)	63.7	66.6	**65.1**	76.5	79.5	**78.0**	56.9
FewRel PubMed [14]	$k = 10$	UIE-PCNN [51]	14.4	45.2	21.9	10.3	19.2	13.5	7.23
		UIE-BERT	10.0	100	18.2	0	100	0	0
		EType+ [58]	10.0	100	18.1	0	100	0	0
		SelfORE	53.7	66.1	59.3	58.8	68.7	63.4	45.4
		PromptORE (RoBERTa)	73.7	73.2	73.5	76.5	77.2	76.9	68.1
		PromptORE (BERT)	77.6	77.2	**77.4**	81.0	81.2	**81.1**	**73.8**

Table 2: Results of PromptORE with different prompts. PromptORE is trained with the exact number of relations k.

Dataset	Prompt	B^3 (F1)	V (F1)	ARI
FewRel	PromptORE ($\mathcal{P}$)	48.8	71.8	43.4
	$\mathcal{P}_\emptyset$	33.8	57.4	28.8
	$\mathcal{P}'_1$	48.9	71.7	44.5
	$\mathcal{P}'_2$	49.4	72.4	46.3
	$\mathcal{P}'_3$	50.5	73.0	47.7
FewRel NYT	PromptORE ($\mathcal{P}$)	65.1	78.0	56.9
	$\mathcal{P}_\emptyset$	51.3	65.7	41.6
	$\mathcal{P}'_1$	65.8	77.8	62.0
	$\mathcal{P}'_2$	61.0	74.8	56.9
	$\mathcal{P}'_3$	65.6	77.7	61.7
FewRel PubMed	PromptORE ($\mathcal{P}$)	77.4	81.1	73.8
	$\mathcal{P}_\emptyset$	62.0	66.2	53.1
	$\mathcal{P}'_1$	76.4	80.0	72.3
	$\mathcal{P}'_2$	76.0	80.0	72.9
	$\mathcal{P}'_3$	77.4	81.1	73.1

Performance on domain specific datasets BERT and more broadly PLMs are usually pretrained on general domain data (e.g. Wikipedia), and we can ask ourselves if that impacts performances on "out of domain" datasets such as FewRel NYT and FewRel PubMed. We can see in Table 1 that **PromptORE** does not see its results plummet. On the contrary, it still outperforms previous SOTA models by a large margin. **SelfORE**, which also relies on BERT embeddings, does not see its performance deteriorate as well, which seems to indicate BERT's ability to encode tokens not seen before.

In general, the results are higher than with FewRel, but that is explained by the fact that the two datasets contain less relations and instances.

Finally, we notice that **Etype+** predicts a single class on FewRel PubMed (as V-measure and ARI touch zero). As we have stated earlier, it is explained by the entity type schema, which is very limited as there are no *Person*, *Organization* or *Location* entities in this dataset.

Does PromptORE really extract relations? The core of **PromptORE** is its prompt $\mathcal{P}$ that is used by the Relation Encoder to embed each instance. However, one can ask if BERT really uses the text of the current instance to predict the missing token (and thus extracts information from the sentence), or if it is only using its internal knowledge, ignoring the current instance context. To answer this question, we propose to create an empty prompt $\mathcal{P}_\emptyset$ where we do not input the current instance text. Its template is defined as:

$$\mathcal{P}_\emptyset(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) = \texttt{"[CLS] } \mathbf{e1} \texttt{ [MASK] } \mathbf{e2}\texttt{.[SEP]"} \tag{2}$$

It is equivalent to $\mathcal{P}$ defined in eq. (1), except that we have removed $\mathbf{S}$.

The results are shown in Table 2. We can see that the performance for all three metrics and three datasets are much lower with $\mathcal{P}_\emptyset$ compared to $\mathcal{P}$, with an average gap of 15% in B^3, 14% in V-measure and 15% in ARI. Therefore, it shows that BERT really benefits from the instance context to extract more precisely the relation between the two entities. It is interesting to remark that even without the instance text,

PromptORE still surpasses **SelfORE**, which clearly indicates that **SelfORE** fails to take full advantage of BERT embeddings.

Alternative Prompts As we have discussed in section 3.1, $\mathcal{P}$ is not necessarily the best prompt, as more relations can be expressed with a noun than with a verb. We computed **PromptORE** performances with three alternative prompts: $\mathcal{P}'_1$, which encourages BERT to predict a noun, $\mathcal{P}'_2$ with the prefix proposed by [35], and $\mathcal{P}'_3$ containing a prefix variant of $\mathcal{P}'_2$. Results are shown in Table 2.

First, we notice that $\mathcal{P}'_1$ provides better results than $\mathcal{P}$ in ARI, but similar performances in V-measure and B^3 for FewRel and FewRel NYT. No improvement is observed with FewRel PubMed. This result is interesting because we showed that fewer relations can be expressed with a verb than with a noun, so we expected a gap in favor of $\mathcal{P}'_1$. For example FewRel relations `instance_of`, `competition_class`, `constellation` or `operating_system` cannot be expressed with a verb but are nonetheless correctly identified with $\mathcal{P}$. It seems that BERT is weakly impacted by the apparent impossibility to predict a meaningful word.

In Table 2, $\mathcal{P}'_2$ and $\mathcal{P}'_3$ achieve higher performances than $\mathcal{P}'_1$ in the majority of the cases, while their only difference with $\mathcal{P}'_1$ is the prefix (*In this sentence* or *We deduce that*). We can also see the impact of prompt's wording: at a first glance both prefixes seem to convey the same idea, but their performances are different. In fact if we replace *deduce* by *conclude* in $\mathcal{P}'_3$ we obtain lower performances (not shown in Table 2).

Finally, we notice that there is no consensus on the best prompt from the four proposed ones: $\mathcal{P}'_3$ is the best for FewRel, $\mathcal{P}'_2$ for FewRel NYT and $\mathcal{P}$ for FewRel PubMed. This highlights the importance to select and fine-tune prompts to maximize BERT's performances, which is indeed a major research area for prompt-based methods [35, 49, 20, 25]. Under our fully unsupervised setting's goal, it is unfeasible to fine-tune the prompt due to the lack of labeled data, therefore we decide to keep our original $\mathcal{P}$ for the sake of fair results.

5.2 Clustering without knowing k

Up to now, **PromptORE** has access to k, the number of different relations. However, as said in section 3.2, the most general setting is when we do not know k. We identified two methods to cluster our data without k: 1. OPTICS, a clustering algorithm based on density, and 2. the Elbow Rule (to compute $\hat{k}$ an estimation of k) with k-means clustering. The results are shown in Table 3.

To give technical details, we use the `scikit-learn` implementation of OPTICS and k-means. To apply the Elbow Rule, we first calculate multiple clusterings by varying the number of clusters. For each of these clusterings we compute the silhouette coefficient. We obtain the blue scatter plot of the Figure 2 for FewRel. As this plot is rough, we approximate it thanks to a ridge regression with a gaussian kernel (orange curve in Figure 2). In our case, this curve has a maximum, it is therefore easy to locate the elbow. We noticed the same curve shape with a maximum for FewRel NYT and FewRel PubMed. Sometimes however, it is possible to observe a growing curve, in which case automatic approaches [27] can be applied to locate the elbow. We detect the elbow at $\hat{k} = 65$ clusters for FewRel. We obtain $\hat{k} = 26$ for FewRel NYT and $\hat{k} = 10$ for FewRel PubMed, that is, values of $\hat{k}$ nearly identical to the real number of relations.

Quantitative Results OPTICS sets $\hat{k}$ at 571 (see Table 3), far from the optimal $k = 80$ for FewRel. It translates into very poor performances compared to **PromptORE** when we know k. We also note that OPTICS is very slow during training ($\sim$ 6h

Table 3: Results of PromptORE using different methods to estimate k. *"Ideal"* represents results when k is provided.

Dataset	Method	$\hat{k}$	B^3 (F1)	V (F1)	ARI
FewRel	*Ideal*	*80*	48.8	**71.8**	**43.4**
	OPTICS	571	10.8	8.5	0
	Elbow	65	**49.5**	71.2	42.2
FewRel NYT	*Ideal*	*25*	**65.1**	**78.0**	**56.9**
	OPTICS	35	33.2	29.3	1.7
	Elbow	26	64.1	77.4	56.2
FewRel PubMed	*Ideal*	*10*	**77.4**	**81.1**	**73.8**
	OPTICS	12	26.8	11.2	0.3
	Elbow	10	**77.4**	**81.1**	**73.8**

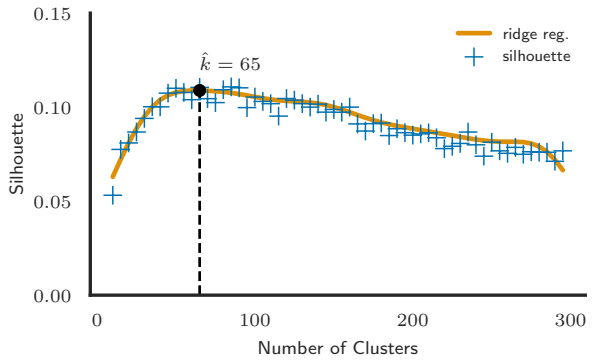

Figure 2: Results of the Elbow Rule on FewRel. Estimated number of relations $\hat{k}$ is equal to 65.

compared to 5min with k-means). On the other side, results are much more satisfactory with the Elbow Rule with a slight decrease in ARI but equivalent performances in B^3 and V-measure. Training time is also much more reasonable with $\sim$ 1h. We make the same conclusion when we look at FewRel NYT and FewRel PubMed.

We can conclude that, at least on our three datasets, the Elbow Rule is effective to find a correct estimation of k.

Finally, it is interesting to see that **PromptORE** with the Elbow Rule widely surpasses previous SOTA approaches (Table 1), with a gap of 15-25% in B^3 and V-measure and 17-30% in ARI. In our opinion, we demonstrate that it is possible to remove the dependency on all hyperparameters (including k) and still achieve state-of-the-art results.

Qualitative Analysis of the Clustering From Table 3, we know that the Elbow Rule finds $\hat{k} = 65$ instead of 80 for FewRel. It means that the clustering cannot be ideal: some clusters must contain multiple relations. The confusion matrix between the true relations and the ones predicted by **PromptORE** is shown in Figure 3. It is obviously not square as the number of clusters $\hat{k}$ is not equal to the number of relations k. We reorganize the axes to find a logical representation of the confusion matrix (as initially there is no link between the relation ids and the cluster ids).

On this confusion matrix, we notice indeed that some clusters are not pure: they

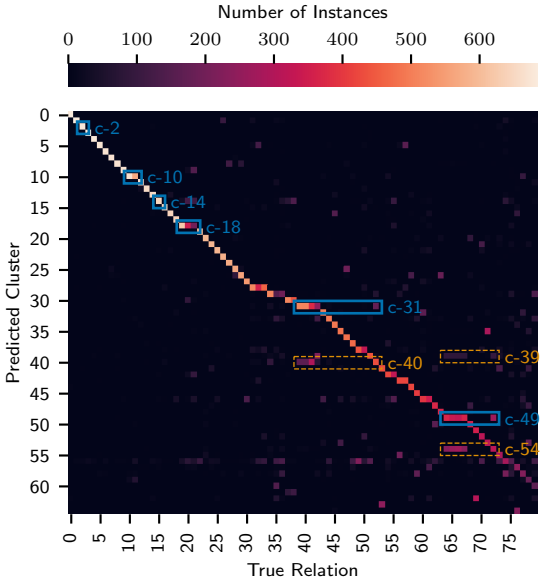

Figure 3: Confusion matrix between the true relations and the clusters found by PromptORE with the Elbow Rule on FewRel. The main relations of some clusters are highlighted.

contain multiple relations (e.g. clusters c-10, c-18, c-28, c-29, c-30, c-31, c-49, or c-54). The main observation is nevertheless that the matrix possesses a clear diagonal, meaning that **PromptORE** is able to effectively distinguish the vast majority of the relations, while training in a fully unsupervised setting.

We also see that clusters seems to be relatively complete: there are seldom clusters sharing the same relations except for clusters c-31 and c-40; c-39, c-49 and c-54.

As some clusters contain multiple relations, we find it interesting to check whether the relations that compose each of these clusters are semantically linked. We randomly sample four clusters that contain multiple relations. Results are shown in Table 4. For these four clusters, we can see that the relations are indeed semantically close within a cluster:

- for cluster c-10 they are linked to language,
- for cluster c-18 to geographical location,
- for cluster c-31 to artistic creation,
- for cluster c-49 to family relationship.

Even though these clusters are not optimal from the FewRel annotation point of view, they are semantically coherent. In fact, we could even argue that cluster c-10 makes more sense than the initial labeling which divided this cluster in two relations.

In conclusion, this qualitative analysis shows that the combination of **PromptORE** with the Elbow Rule efficiently discovers semantically consistent clusters, which are very close to true relations.

5.3 Analysis of $\mathcal{P}$ Prompt Predictions

In section 5.1, we found a very little performance difference between $\mathcal{P}$ and $\mathcal{P}'_1$, while the majority of relations cannot be written using a verb. To go further, it would be

Table 4: Relation distributions that compose four randomly sampled impure clusters.

Cluster	True Relations
c-10	language of film or TV show, language of work or name
c-18	mountain range, located in physical feature, located in or next to body of water
c-31	screenwriter, director, after a work by, characters, composer
c-49	sibling, father, child, mother, spouse

interesting to check which tokens/verbs seem to describe best the clusters identified by **PromptORE** with $\mathcal{P}$. To do that, we use the relation embeddings (computed by our Relation Encoder) and predict the masked tokens (represented by `[MASK]`) with the MLM head of BERT. By iterating for every instance located in one cluster, we can find the most frequent tokens that seem to describe it. We apply this method on the clusters identified by **PromptORE** and Elbow Rule (see Figure 3). The results on three selected clusters are displayed in Table 5.

We observe that in most cases, the predicted names are not clear enough to qualify the relation corresponding to the cluster. Nevertheless, the names give clues to identify the general theme of the relation (*married* indicates a family centered relation, *borders* and *surrounds* a geographic topic).

Table 5 shows however the major limitation of $\mathcal{P}$: when we look at cluster c-49, `spouse` is represented by *married*, but `sibling`, `father`, `child` and `mother` relations are not brought to light with the predicted names. Indeed, these four relations cannot be written using a single verb; by not finding satisfactory names, BERT defaulted into predicting punctuation tokens. We reach the same conclusion for cluster c-14, where BERT also predicted punctuation tokens.

Finally, this observation gives us interesting insights on the behavior of our Relation Encoder. Intuitively, we could think that **PromptORE** would have poor results with relations that cannot be written using a verb. On the contrary, we found that results were close between $\mathcal{P}$ and $\mathcal{P}'_1$ (Table 2). At the same time, we notice that cluster c-2 (Table 5) is very pure, yet its most predicted name is *","* (a token that is furthermore shared among the two other clusters of Table 5). In our opinion, it indicates that BERT is able to encode a very expressive embedding of the current relation instance that allows a precise clustering, but that cannot be translated into real words. This is supported by the fact that **PromptORE** identified three different clusters with punctuation as their most frequent tokens (Table 5). It comforts us in the idea that complex prompts are not required to effectively represent a significant number of relations. It does not undermine however the importance of prompt-tuning: as showed in Table 2, prompts have an impact on model performances.

6 Conclusion and Future Work

In this work, we introduce **PromptORE**, an unsupervised RE model. Our proposed approach leverages and adapts the novel prompt-tuning paradigm. Experiments on one general and two domain specific datasets show that **PromptORE** surpasses previous state-of-the-art methods, while being simpler and not needing hyperparameter tuning.

Table 5: Most frequent predicted tokens for three different clusters identified by PromptORE for FewRel. True relations composing these clusters are also displayed.

Cluster	Predicted tokens	True relations
c-2	, for supports	99%: platform *other*
c-14	: borders , surrounds	79%: contains administrative territory 5%: located in administrative entity 3%: located in physical feature *other*
c-49	, . married	25%: sibling 21%: father 19%: child 18%: mother 17%: spouse

On a secondary note, finding descriptive names for the clusters is still an open question.

In the future, we plan to explore other clustering approaches with a focus on Deep Clustering methods (e.g. [44]), and Hierarchical Clustering models that leverage the hierarchical nature of relations. We further envision to *close the loop* of knowledge extraction, that is, benefiting from **PromptORE**'s ability to extract relations in order to build a knowledge graph that can be used to further improve **PromptORE**'s predictions.

Acknowledgments

This work is supported by Alteca and the French Association for Research and Technology (ANRT) under CIFRE PhD fellowship n°2021/0851.

References

[1] Mihael Ankerst, Markus M. Breunig, Hans Peter Kriegel, and Jörg Sander. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, (2):49–60, 6 1999.

[2] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chain. In *Proceedings of the 1st International Conf. on Language Resources and Evaluation Workshop on Linguistics Coreference*, page 563–566, 1998.

[3] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conf. on Artificial Intelligence*, pages 2670–2676, Hyderabad, India, 2007. Morgan Kaufmann Publishers Inc.

[4] David M Blei, Andrew Y Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, (4-5):993–1022, 2003.

[5] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In *Proceedings of the ACM Web Conf. 2022*, Lyon, France, 4 2022. ACM.

[6] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, page 112948, 3 2020.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 539–546, San Diego, CA, USA, 2005. IEEE Computer Society.

[8] Lei Cui, Furu Wei, and Ming Zhou. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 407–413, Melbourne, Australia, 2018. ACL.

[9] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conf. of the NAACL: Human Language Technologies*, pages 4171–4186, Stroudsburg, PA, USA, 2019. ACL.

[10] Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest. Unsupervised Open Relation Extraction. In *The Semantic Web: ESWC 2017 Satellite Events*, pages 12–16, Cham, 1 2017. Springer.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conf. on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, United States, 1996. AAAI Press.

[12] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, (5814):972–976, 2 2007.

[13] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conf. on Natural Language Processing*, pages 3816–3830, Online, 12 2021. ACL.

[14] Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing and 9th International Joint Conf. on Natural Language Processing*, pages 6250–6255, Hong Kong, China, 2019. ACL.

[15] Jiaying Gong and Hoda Eldardiry. Prompt-based Zero-shot Relation Classification with Semantic Knowledge Augmentation, 12 2021.

[16] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A Survey on Knowledge Graph-Based Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–17, 10 2020.

[17] Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction. In *Proceedings of the 1st Conf. of the Asia-Pacific Chapter of the ACL and the 10th International Joint Conf. on Natural Language Processing*, pages 745–758, Suzhou, China, 2020. ACL.

[18] Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing: System Demonstrations*, pages 169–174, Hong Kong, China, 2019. ACL.

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

[20] Adi Haviv, Jonathan Berant, and Amir Globerson. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conf. of the European Chapter of the ACL*, pages 3618–3623, Online, 2021. ACL.

[21] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conf. on Empirical Methods in Natural Language Processing*, pages 3673–3682, Online, 2020. ACL.

[22] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. Knowledge graph embedding based question answering. In *Proceedings of the 12th ACM International Conf. on Web Search and Data Mining*, pages 105–113, Melbourne, Australia, 1 2019. ACM, Inc.

[23] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, (1):193–218, 12 1985.

[24] Shaoxiong Ji, Shirui Pan, Erik Cambria, Senior Member, Pekka Marttinen, Philip S. Yu, and Life Fellow. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1 – 21, 2021.

[25] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the ACL*, pages 423–438, 2020.

[26] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conf. on Learning Representations*, Banff, Canada, 12 2014. International Conf. on Learning Representations.

[27] Pavel V Kolesnichenko, Qianhui Zhang, Changxi Zheng, Michael S Fuhrer, and Jeffrey A Davis. Multidimensional analysis of excitonic spectra of monolayers of tungsten disulphide: toward computer-aided identification of structural and environmental perturbations of 2D materials. *Machine Learning: Science and Technology*, (2):025021, 3 2021.

[28] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A Joint Neural Model for Information Extraction with Global Features. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 7999–8009, Online, 7 2020. ACL.

[29] Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. Element intervention for open relation extraction. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conf. on Natural Language Processing*, pages 4683–4693, Online, 6 2021. ACL.

[30] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, 7 2021.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 7 2019.

[32] S. P. Lloyd. Least squares quantization in PCM. *Technical Report RR-5497*, 1957.

[33] Renze Lou, Fan Zhang, Xiaowei Zhou, Yutong Wang, Minghui Wu, and Lin Sun. A Unified Representation Learning Strategy for Open Relation Extraction with Ranked List Loss. In *Proceedings of the 20th China National Conf. on Computational Linguistics*, pages 1096–1108, Huhhot, China, 2021. Chinese Information Processing Society of China.

[34] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conf. of the NAACL: Human Language Technologies*, pages 3036–3046, Minneapolis, Minnesota, 2019. ACL.

[35] Bo Lv, Li Jin, Yanan Zhang, Hao Wang, Xiaoyu Li, and Zhi Guo. Commonsense Knowledge-Aware Prompt Tuning for Few-Shot NOTA Relation Classification. *Applied Sciences*, (4):2185, 2 2022.

[36] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symp. on mathematical statistics and probability*, pages 281–297, Berkeley, California, United States, 1967. University of California Press.

[37] Diego Marcheggiani and Ivan Titov. Discrete-State Variational Autoencoders for Joint Discovery and Factorization of Relations. *Transactions of the ACL*, pages 231–244, 12 2016.

[38] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th International Joint Conf. on Natural Language*, pages 1003–1011, Suntec, Singapore, 2009. ACL.

[39] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conf. on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, 2014. ACL.

[40] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True Few-Shot Learning with Language Models, 5 2021.

[41] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conf. of the NAACL: Human Language Technologies*, pages 2227–2237, New Orleans, Louisiana, United States, 2 2018. ACL.

[42] Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. A Two-phase Prototypical Network Model for Incremental Few-shot Relation Classification. In *Proceedings of the 28th International Conf. on Computational Linguistics*, pages 1618–1629, Barcelona, Spain, 1 2020. ACL.

[43] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, number PART 3, pages 148–163, Berlin, Heidelberg, 2010. Springer.

[44] Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. DeepDPM: Deep Clustering With an Unknown Number of Clusters, 3 2022.

[45] Andrew Rosenberg and Julia Hirschberg. V-Measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, Prague, Czech Republic, 2007. ACL.

[46] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, (C):53–65, 11 1987.

[47] Arpita Roy, Youngja Park, Taesung Lee, and Shimei Pan. Supervising unsupervised open information extraction models. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing and 9th International Joint Conf. on Natural Language Processing*, pages 728–737, Hong Kong, China, 2019. ACL.

[48] Swarnadeep Saha and Mausam. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conf. on Computational Linguistics*, pages 2288–2299, Santa Fe, New Mexico, USA, 2018. ACL.

[49] Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, and Sameer Singh. AUTOPROMPT: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conf. on Empirical Methods in Natural Language Processing*, pages 4222–4235, Online, 10 2020. ACL.

[50] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, (1):30–34, 1 1973.

[51] Étienne Simon, Vincent Guigue, and Benjamin Piwowarski. Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 1378–1387, Florence, Italy, 2019. ACL.

[52] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4078–4088. Neural information processing systems foundation, 3 2017.

[53] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 2895–2905, Florence, Italy, 2019. ACL.

[54] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised open information extraction. In *Proceedings of the 2018 Conf. of the NAACL: Human Language Technologies*, pages 885–895, New Orleans, Louisiana, United States, 2018. ACL.

[55] Douglas Steinley. Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, (3):386–396, 9 2004.

[56] Jiejun Tan, Wenbin Hu, and WeiWei Liu. EPPAC: Entity Pre-typing Relation Classification with Prompt Answer Centralizing, 3 2022.

[57] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, (4):267–276, 12 1953.

[58] Thy Thy Tran, Phong Le, and Sophia Ananiadou. Revisiting Unsupervised Relation Extraction. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 7498–7505, Online, 7 2020. ACL.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems 30*, pages 5999–6009. Neural information processing systems foundation, 6 2017.

[60] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing and 9th International Joint Conf. on Natural Language Processing*, pages 5784–5789, Hong Kong, China, 2019. ACL.

[61] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M. Robertson. Ranked list loss for deep metric learning. In *Proceedings of the 2019 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 5202–5211, Long Beach, CA, United States, 3 2019. IEEE Computer Society.

[62] Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. Open relation extraction: Relational knowledge transfer

from supervised data to unsupervised data. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing and 9th International Joint Conf. on Natural Language Processing*, pages 219–228, Hong Kong, China, 2019. ACL.

[63] Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. Structured relation discovery using generative models. In *Proceedings of the 2011 Conf. on Empirical Methods in Natural Language Processing*, pages 1456–1466, Edinburgh, Scotland, UK, 2011. ACL.

[64] Limin Yao, Sebastian Riedel, and Andrew McCallum. Unsupervised relation discovery with sense disambiguation. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 712–720, Jeju Island, Korea, 2012. ACL.

[65] Chenhan Yuan and Hoda Eldardiry. Unsupervised Relation Extraction: A Variational Autoencoder Approach. In *Proceedings of the 2021 Conf. on Empirical Methods in Natural Language Processing*, pages 1929–1938, Stroudsburg, PA, USA, 2021. ACL.

[66] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conf. on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, 2015. ACL.

[67] Kai Zhang, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. Open Hierarchical Relation Extraction. In *Proceedings of the 2021 Conf. of the NAACL: Human Language Technologies*, pages 5682–5693, Online, 6 2021. ACL.

[68] Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *Proceedings of the 2019 Conf. of the NAACL: Human Language Technologies*, pages 3016–3025, Minneapolis, Minnesota, USA, 2019. ACL.

[69] Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. De-biasing Distantly Supervised Named Entity Recognition via Causal Intervention. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conf. on Natural Language Processing*, pages 4803–4813, Online, 6 2021. ACL.

[70] Wenkai Zhang, Hongyu Lin, Xianpei Han, Le Sun, Huidan Liu, Zhicheng Wei, and Nicholas Jing Yuan. Denoising Distantly Supervised Named Entity Recognition via a Hypergeometric Probabilistic Model. *Proceedings of the 35th AAAI Conf. on Artificial Intelligence*, (16):14481–14488, 6 2021.

[71] Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. A Relation-Oriented Clustering Method for Open Relation Extraction. In *Proceedings of the 2021 Conf. on Empirical Methods in Natural Language Processing*, pages 9707–9718, Punta Cana, Dominican Republic, 2021. ACL.

[72] Shun Zheng, Xu Han, Yankai Lin, Peilin Yu, Lu Chen, Ling Huang, Zhiyuan Liu, and Wei Xu. DIAG-NRE: A neural pattern diagnosis framework for distantly supervised neural relation extraction. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 1419–1429, Florence, Italy, 2019. ACL.

[73] Zexuan Zhong and Danqi Chen. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Annual Conf. of the NAACL*, pages 50–61, Online, 2021. ACL.