PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction

To cite this version:

HAL Id: hal-03858264
https://hal.science/hal-03858264v2
Submitted on 23 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License
PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction

Pierre-Yves Genest\textsuperscript{1, 2}, Pierre-Edouard Portier\textsuperscript{2}, Előd Egyed-Zsigmond\textsuperscript{2}, and Laurent-Walter Goix\textsuperscript{1}

\textsuperscript{1}Alteca, 88 Boulevard des Belges, 69006 Lyon, France
\textsuperscript{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 \textit{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\textsuperscript{3}, V-measure and ARI.

Qualitative analysis also indicates PromptORE’s ability to identify semantically coherent clusters that are very close to true relations\footnote{Source code is available at \url{https://github.com/alteca/PromptORE}}.

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

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, 32, 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 hy-
perparameter 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 $(e_1, \text{relation}, e_2)$ 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 [15, 17]. Recent approaches follow a two-step process [17, 28, 31, 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 [54], 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. [31] 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. [12] 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 \([35,36,16,53]\). Efforts are focused in optimizing the prompt \(P\) 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 \([13,35]\)).

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. \([25]\) 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 \([63,64,37]\). Elsayah et al. \([47]\) 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 \([31]\) 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. \([52]\) 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 \([42]\). 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 \([37,65,58,21,62,53]\), learning rate, regularization \([51]\), entity types \([53]\), early-stopping \([55]\), 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 \(e_1\) and \(e_2\) present in the same sentence\(^2\). 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 \(D\) (see Figure \(1\)) containing instances with the following properties:

- An instance is described with a triple \((S, e_1, e_2)\), where \(S = [t_0, ..., t_{s-1}]\) is the instance text composed of tokens\(^3\) and \(e_1 = [t_{\text{start}(e_1)}, ..., t_{\text{end}(e_1)}]\) and \(e_2 = [t_{\text{start}(e_2)}, ..., t_{\text{end}(e_2)}]\) are two entities identified by their indexes in \(S\).
- We suppose that \(e_1\) and \(e_2\) have already been extracted (but not typed).
- In the instance text \(S\), \(e_1\) and \(e_2\) are linked by a binary relation \(r\). As previous approaches, we do not consider the case where there is no relation between \(e_1\) and \(e_2\).
- We do not have access to any relation label during training and hyperparameter tuning.

\(^2\)As previous works, we focus on sentence RE, even though we are aware that some relations may be missed.

\(^3\)Token as defined by BERT \([9]\): punctuation, word or part of word.
Relation Clustering

Dataset \( \mathcal{D} \)
1. Queen Elizabeth II was married to Prince Philip.
2. Michelle Obama, Barack Obama’s wife, was born in Chicago.
   n. Like the Everest, Mount Annapurna is located in the Himalayas.

1) Relation Encoder

Prompt generation
\[ \mathcal{P} \] (CLS) Queen Elizabeth II was married to Prince Philip.
Queen Elizabeth II [MASK] Prince Philip (SEP)

Prompt embedding
[CLS] Queen ... Philip ... Queen ... [MASK] ... Philip ... [SEP]

BERT

2) Relation Clustering

Cluster c-49

Cluster c-18

Cluster c-2

Figure 1: Overview of PromptORE.


\( \mathcal{R} \) 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.

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.

3.1 Relation Encoder

This module aims to compute a vector representation (or relation embedding) of the relation expressed between \( e_1 \) and \( e_2 \) 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 [19, 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:

- [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 \( h = \text{BERT}(S, [\text{MASK}]) \) as the BERT-embedding of the token [MASK], in the context of \( S \) (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 \( a = \text{One} \), "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 [16, 15, 35, 50]. 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 \( \mathcal{P}(S, e_1, e_2) = \"[CLS]\ S \text{ e1 e2 \}[SEP]\" \).

2. Identify a set of label tokens \( \mathcal{L} \) that represents each relation.

3. Predict the [MASK] embedding \( \mathcal{L} \) that represents each relation.

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:

\[
\mathcal{P}(S, e_1, e_2) = \"[CLS]\ S \text{ e1 [MASK] e2 [SEP]}\" \tag{1}
\]
with $S$ the instance text, $e_1$ (resp. $e_2$) the text of the first (resp. second) entity. For example, given the sentence $S = \text{Queen Elizabeth II was married to Prince Philip.}$, and entities $e_1 = \text{Queen Elizabeth II}$ and $e_2 = \text{Prince Philip}$, the prompt is:

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

As we remove $L$, we decide to use the $[\text{MASK}]$ BERT embedding as our relation embedding. Thus, the Relation Encoder process is the following (as shown in Figure 4):

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

**Alternative Prompts** If we analyze $P$, we can see that BERT will likely fill the $[\text{MASK}]$ 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 9892 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:

$$P'_1(S, e_1, e_2) = [\text{CLS}] S e_1 \text{ is the } [\text{MASK}] \text{ of } e_2. [\text{SEP}]$$  
$$P'_2(S, e_1, e_2) = [\text{CLS}] S \text{ In this sentence, } e_1 \text{ is the } [\text{MASK}] \text{ of } e_2. [\text{SEP}]$$  
$$P'_3(S, e_1, e_2) = [\text{CLS}] S \text{ We deduce that } e_1 \text{ is the } [\text{MASK}] \text{ of } e_2. [\text{SEP}]$$  
$$P'_4$$ 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 $P$, 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] [7], we cluster the relation embeddings computed on the entire dataset $D$ 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 [17], OPTICS [1] or Affinity Propagation [12]. Nevertheless, most cannot be applied in our case: DBSCAN and HAC need other hyperparameters (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 [26] as our internal metric 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 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]. 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]. 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 [17]. The text comes from PubMed, a database of biomedical literature. It is also automatically annotated (this time with the UMLS knowl-

---

5 A metric that does not rely on external data such as labels.

6 Data can be downloaded from: [https://github.com/thunlp/FewRel](https://github.com/thunlp/FewRel)
edge 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, 61], 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+ [38] and UIE-PCNN [61]).

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.

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 10% 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, 59]. 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, 07, 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 20-50%). 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, 07]. 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.

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

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

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:

$$\mathcal{P}_\emptyset(S, e_1, e_2) = "[CLS] e_1 [MASK] e_2 [SEP]"$$

(2)

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

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 [53], and $\mathcal{P}'_3$ containing a prefix variant of $\mathcal{P}'_2$. Results are shown in Table 2.

![](image-url)
Table 3: Results of PromptORE using different methods to estimate $k$. "Ideal" represents results when $k$ is provided.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>$k$</th>
<th>$B^3$ (F1)</th>
<th>V (F1)</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FewRel</td>
<td>Ideal</td>
<td>80</td>
<td>48.8</td>
<td>71.8</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>OPTICS</td>
<td>571</td>
<td>10.8</td>
<td>8.5</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Elbow</td>
<td>65</td>
<td>49.5</td>
<td>71.2</td>
<td>42.2</td>
</tr>
<tr>
<td>FewRel</td>
<td>Ideal</td>
<td>25</td>
<td>65.1</td>
<td>78.0</td>
<td>56.9</td>
</tr>
<tr>
<td>NYT</td>
<td>OPTICS</td>
<td>35</td>
<td>33.2</td>
<td>29.3</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Elbow</td>
<td>26</td>
<td>64.1</td>
<td>77.4</td>
<td>56.2</td>
</tr>
<tr>
<td>FewRel</td>
<td>Ideal</td>
<td>10</td>
<td>77.4</td>
<td>81.1</td>
<td>73.8</td>
</tr>
<tr>
<td>PubMed</td>
<td>OPTICS</td>
<td>12</td>
<td>26.8</td>
<td>11.2</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Elbow</td>
<td>10</td>
<td>77.4</td>
<td>81.1</td>
<td>73.8</td>
</tr>
</tbody>
</table>

First, we notice that $P'_1$ provides better results than $P$ 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 $P'_1$. For example, FewRel relations instance_of, competition_class, constellation or operating_system cannot be expressed with a verb but are nonetheless correctly identified with $P$. It seems that BERT is weakly impacted by the apparent impossibility to predict a meaningful word.

In Table 2, $P'_2$ and $P'_3$ achieve higher performances than $P'_1$ in the majority of the cases, while their only difference with $P'_1$ 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 $P'_3$ 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: $P'_1$ is the best for FewRel, $P'_2$ for FewRel NYT and $P$ 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. 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 $P$ 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 5.2 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 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 (~6h 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 ~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 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 seem 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 5.2, 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 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 \([\text{MASK}]\)) 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 4.

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 (\textit{married} indicates a family centered relation, \textit{borders} and \textit{surrounds} a geographic topic).

Table 4 shows however the major limitation of \( \mathcal{P} \): when we look at cluster c-49, \textit{spouse} is represented by \textit{married}, but \textit{sibling}, \textit{father}, \textit{child} and \textit{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.
Table 5: Most frequent predicted tokens for three different clusters identified by PromptORE for FewRel. True relations composing these clusters are also displayed.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Predicted tokens</th>
<th>True relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>c-2</td>
<td>for supports</td>
<td>99%: platform other</td>
</tr>
<tr>
<td>c-14</td>
<td>borders, surrounds</td>
<td>79%: contains administrative territory 5%: located in administrative entity 3%: located in physical feature other</td>
</tr>
<tr>
<td>c-49</td>
<td>married</td>
<td>25%: sibling 21%: father 19%: child 18%: mother 17%: spouse</td>
</tr>
</tbody>
</table>

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. 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


[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.


[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.


[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, Sun-tec, Singapore, 2009. 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 728–737, Hong Kong, China, 2019. ACL.


[58] David Wadden, Ulme Wennberg, Yi Luan, and Hanneh 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.


