# High Recall Open IE for Relation Discovery

**Hady Elsahar, Christophe Gravier, Frederique Laforest**
Université de Lyon, Laboratoire Hubert Curien, Saint-Étienne, France
hady.elsahar@univ-st-etienne.fr
christophe.gravier@univ-st-etienne.fr
frederique.laforest@univ-st-etienne.fr

## Abstract

Relation Discovery discovers predicates (relation types) from a text corpus relying on the co-occurrence of two named entities in the same sentence. This is a very narrowing constraint: it represents only a small fraction of all relation mentions in practice. In this paper we propose a high recall approach for predicate extraction which enables covering up to 16 times more sentences in a large corpus. Comparison against OpenIE systems shows that our proposed approach achieves 28% improvement over the highest recall OpenIE system and 6% improvement in precision over the same system.

## 1 Introduction

The recent years have shown a large number of knowledge bases such as YAGO (Suchanek et al., 2007), Wikidata (Vrandečić and Krötzsch, 2014) and Freebase (Bollacker et al., 2008). These knowledge bases contain information about world entities (e.g. countries, people...) using a set of predefined predicates (e.g. birth place, profession...) that comes from a fixed ontology. The number of predicates can vary according to the KB ontology. For example there are 6,1047 DBpedia unique predicates compared to only 2,569 in Wikidata. This has led to an emergence of unsupervised approaches for relation extraction which can scale to open relations that are not predefined in a KB ontology.

### 1.1 Open Information Extraction

Open information extraction (Open IE) systems extract linguistic relations in the form of tuples from text through a single data-driven pass over a large text corpus. Many Open IE systems have been proposed in the literature, some of them are based on patterns over shallow syntactic representations such as TEXTRUNNER (Banko et al., 2007) and REVERB (Fader et al., 2011), pattern learning in OLLIE (Mausam et al., 2012), Tree Kernels (Xu et al., 2013) or logic inference in STANFORD OPEN IE (Angeli et al., 2015).
Open IE has demonstrated an ability to scale to a non-predefined set of target predicates over a large corpus. However extracting new predicates (relation types) using Open IE systems and merging to existing knowledge bases is not a straightforward process, as the output of Open IE systems contains redundant facts with different lexical forms e.g. *(David Bowie, was born in, London)* and *(David bowie, place of birth, London)*.

### 1.2 Relation Discovery and Clustering

Relation clustering and relation discovery techniques try to alleviate this problem by grouping surface forms between each pair of entities in a large corpus of text. A large body of work has been done in that direction, through: clustering of OpenIE extractions (Mohamed et al., 2011; Nakashole et al., 2012a,b), topic modeling (Yao et al., 2011, 2012), matrix factorization (Takamatsu et al., 2011) and variational autoencoders (Marcheggiani and Titov, 2016).
These approaches are successful to group and discover relation types from a large text corpus for the aim of later on adding them as knowledge base predicates.

### 1.3 Relation Discovery with a Single Entity Mention

Previously described relation discovery techniques discover relations between a detected pair of named entities. They usually use a preprocessing step to select only sentences with the mention of a pair of named entities (Figure 1 ex-



*Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 228–233,
Taipei, Taiwan, November 27 – December 1, 2017 

ample 1). This step skips many sentences in which only one entity is detected. These sentences potentially contain important predicates that can be extracted and added to a KB ontology.
Figure 1 illustrates different examples of these sentences, such as: When the object is not mentioned (example 2), Questions where the object is not mentioned (example 3) or when one of the entities is hard to detect because of coreferencing or errors in NER tagging (example 4). By analysing

1. The ***official currency*** of the **U.K.** is the **Pound sterling.**
2. The **U.K.** ***official currency*** is down 16 percent since June 23.
3. What is the ***official currency*** of **U.K.** ?
4. .. which is considered the ***official currency*** of **U.K.**

Figure 1: Examples of textual representations mentioning the predicate "official currency".

**630K documents** from the NYT corpus (Sandhaus, 2008) as illustrated in Figure 2, the number of sentences with two 2 detected named entities is only **1.8M sentences**. Meanwhile, there are almost **30M sentences** with one entity (16 times more), which can be explored for predicate mentions. As the set of covered sentences is limited, so is the number of possibly discovered predicates.
In this paper we propose a predicate-centric method to extract relation types from such sentences while relying on only one entity mention. For relation clustering, we leverage various features from relations, including linguistic and semantic features, and pre-trained word embeddings. We explore various ways of re-weighting and fusing these features for enhancing the clustering results. Our predicate-centric method achieves 28% enhancement in recall over the top Open IE system and with a very comparable precision scores over an OpenIE benchmark (Stanovsky and Dagan, 2016). It demonstrates its superiority for the discovery of relation types.

## 2 Our Approach

### 2.1 Extraction of Predicates

Banko et.al (Banko and Etzioni, 2008) show that the majority of relations in free text can be represented using a certain type of Part of Speech (POS) patterns (e.g. "VB", "VB IN", "NN IN"). Additionally Riedel et al. (Riedel et al., 2013) propose the Universal Schemas model in which the lexicalized dependency path between two named entities in the same sentence is used to represented the relation in between. We follow a similar approach to extract lexical forms of predicates in sentences and connect them to named entities in the sentences.

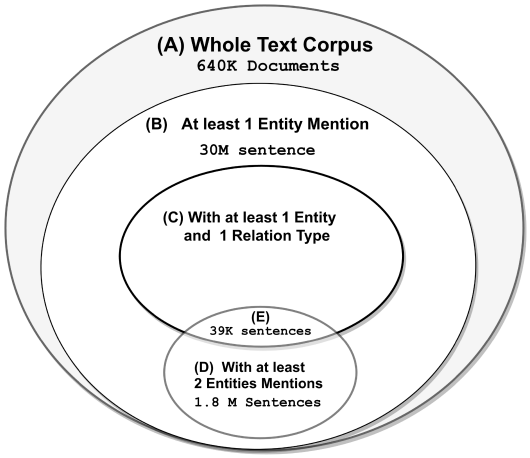


Figure 2: Distribution of sentences in the NYT corpus (A), which have: (B) at least 1 entity mention, (C) at least 1 entity and a predicate attached to it, (D) at least 2 entities mentions, (E) at least 2 entities and a relation in between in Freebase.

First to expand the set of predicate patterns proposed by Banko et al., we collect labels and aliases for 2,405 Wikidata (Vrandečić and Krötzsch, 2014) predicates, align them with sentences from Wikipedia, and run the CoreNLP POS tagger (Manning et al., 2014) on them. This results in a set of 212 unique patterns $POS = \{pos_i, ..., pos_n\}$ [1].
Second, for each sentence in the corpus we do the following:

(i) extract the linguistic surface forms of predicate candidates $P_c$ by matching the POS tagging of the sentence with the set of POS patterns $POS$.

(ii) extract candidate named entities $E_c$ using the CoreNLP NER tagger (Manning et al., 2014).

(iii) extract the lexicalized dependency path $dp_i$ and its direction between every named entity $e_i \in E_c$ and candidate relation predicates $p_i \in P_c$ (if exist). The direction of the dependency path highly correlates with the entity

[1] http://bit.ly/2obhbyF

being subject or object of the candidate predicate (Roth and Lapata, 2016).

The result of this process is a set of extractions $Ext = \{(p_i, e_i, dp_i)...(p_n, e_n, dp_n)\}$, in which a predicate $p_i$ is connected to a named entitiy $e_i$ through a directed dependency paths $dp_i$. We ignore all the candidate predicates that are not connected to a named entity though a dependency path. The confidence for each extraction is calculated according to the rank of its dependency path $dp_i$ and its POS pattern.

### 2.2 Predicates Representation and Clustering

For each predicate in $Ext$, there are predicates though having different surface forms, express the same semantic relationship (e.g. "was born in", "birth place"). Following (Mohamed et al., 2011), we treat predicates with the same surface form as one input to the clustering approach. A feature representation vector for each unique predicate is built from multiple sentences across the text corpus. In the literature, this approach is referred to as the macro scenario, in contrast to the micro scenario (Yao et al., 2011; Marcheggiani and Titov, 2016) where every sentence in the corpus is treated individually. The input to the clustering process in the macro scenario is very small in comparison to the micro scenario, which makes the macro scenario more scalable.

For each unique predicate $p_i \in P$ we built a feature vector that consists of the following set of features:

1. Sum of TF-IDF re-weighted word embeddings for each word in $p_i$.
2. Count vector of each entity appearing as subject and as an object to $p_i$
3. Count vector of entity types appearing as subject and as an object to $p_i$
4. Count vector of each unique dependency path $p_i$ that extracted $p_i$
5. The POS pattern of $p_i$ encoded as a vector containing counts of each POS tag.

The previous features are not equally dense – concatenating all of them as a single feature vector for each relation is expected to skew the clustering algorithm. In supervised relation extraction, this is not an issue as the learning algorithm is expected to do feature selection automatically using training data. Here, it is not the case. In order to circumvent the sparse features bias, we apply individual feature reduction of the sparse features before merging them to the rest of the feature vectors. For feature reduction, we use Principal Component Analysis (PCA) (Jolliffe, 2011). Once this reduction is applied, we apply a K-Means clustering (Hartigan and Wong, 1979) algorithm over the relations feature vectors in order to group relations into $k$ clusters.

## 3 Experiments and Evaluation

### 3.1 Predicates Extraction

In this section we demonstrate the effectiveness of using the proposed predicate-centric approach for relation discovery. For that we use a large scale dataset that was used for benchmarking Open IE (Stanovsky and Dagan, 2016). The dataset is comprised of 10,359 Open IE gold standard extractions over 3,200 sentences. Extractions are evaluated against the gold standard using a matching function between the extracted predicate and candidate predicates from Open IE systems. Extracted predicates that do not exist in the gold standard are calculated as false positives. We compare our predicate extraction method with a set of 6 Open IE systems, which are: REVERB, OLLIE, STANFORD-OPENIE, CLAUSIE (Corro and Gemulla, 2013), OPENIE4.0 an extension of SRL-based IE (Christensen et al., 2011) and noun phrase processing (Pal and Mausam, 2016) , and PROPS (Stanovsky et al., 2016).

Figure 3 shows that our proposed approach scores the highest recall amongst all the Open IE systems with 89% of predicates being extracted, achieving 28% improvement over CLAUSIE, the Open IE system with the highest recall and 6% improvement in precision over the same system. This shows that our approach is more useful when the target application is relation discovery, as it is able to extract predicates in the long tail with comparable precision, as shown in Figure 4. Table 2 shows a set of example sentences in the evaluation dataset in which none of the existing Open Information Extraction systems where able to extract, while they are correctly extracted by our approach.

### 3.2 Quality of Relation Clustering

To the best of our knowledge, the literature does not provide datasets for evaluating Relation Dis-

| Sentence | Target predicate | Predicate-Centric Extraction |
|---|---|---|
| Nicephorus Xiphias , who had conquered the old Bulgarian capitals. | conquered | conquered $\rightarrow$ *dobj* $\rightarrow$ *MISC* |
| Muncy Creek then turns northeast , crossing Pennsylvania Route 405 | crossing | crossing $\rightarrow$ *dobj* $\rightarrow$ *LOCATION* |
| This was replaced by a Town Hall | replaced by | replaced by $\rightarrow$ *nmod* $\rightarrow$ *LOCATION* |
| Starting in 2009 , Akita began experiencing ... | Starting in | Starting in $\rightarrow$ *nmod* $\rightarrow$ *DATE* |

Table 1: Example of sentences where all OpenIE systems failed to extract target relations, and their corresponding Predicate-Centric extractions.

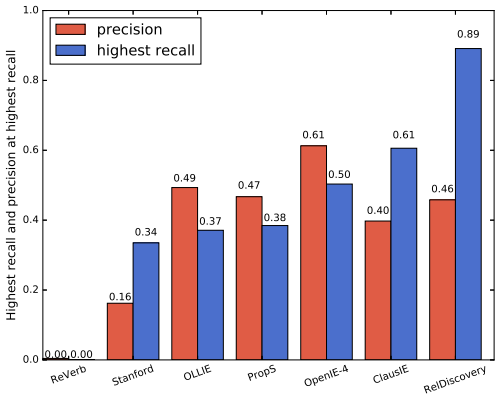


Figure 3: Maximum recall of top Open IE systems and their corresponding precisions in comparison with our approach **RelDiscovery** on (Stanovsky and Dagan, 2016) evaluation dataset.

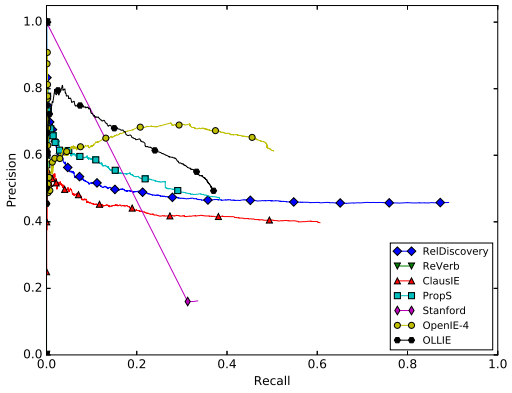


Figure 4: Precision and recall curve of our relation discovery method **RelDiscovery** with different OpenIE systems.

covery methods on the macro scenario. So we use GOOGLE-RE[2], a high quality dataset, that consists of sentences manually annotated with triples from Freebase (Bollacker et al., 2008). The dataset consists of 34,741 labeled sentences, for 5 Freebase relations: "institution", "place of birth", "place of death", "date of birth" and "education degree". We run our predicate extraction approach on the dataset and manually label the most frequent 2K extracted relations into 6 classes: the 5 target semantic relations in GOOGLE-RE and an additional class "OTHER" for other relations. We then divide them to 80-20% test-validation splits. For feature building, we use word2vec pre-trained word embeddings (Mikolov et al., 2013). We tune the PCA using the validation dataset. Results in Table 2 show that the re-weighting of Word embedding using TF-IDF had a significant improvement over only summing word embeddings. This opens the door for exploring more common unsupervised representations for short texts. Additionally, individual feature reduction on the sparse features has significantly enhanced the pairwise F1 score of the clustering algorithm.

[2]http://bit.ly/2oyGBcZ

| Em-Ft | wEm-Ft | wEm-Ft-PCA | ALL |
|---|---|---|---|
| 0.41 | 0.50 | 0.55 | **0.58** |

Table 2: pairwise F1 scores using word embeddings and sparse features (Em-Ft), after re-weighting word embeddings (wEm-Ft), after doing feature reduction (wEm-Ft-PCA), and combining all features (ALL).

## 4 Conclusion

We introduce a high recall approach for predicate extraction. It covers up to 16 times more sentences in a large corpus. Our approach is predicate-centric and learns surface patterns to directly extract lexical forms representing predicates and attach them to named entities. Evaluation on an OpenIE benchmark show that our system was able to achieve a significantly high recall (89%) with 28% improvement over the CLAUSIE, the Open IE system with the highest recall. It shows also a with very comparable precision with the rest of the OpenIE systems. Additionally, we introduce a baseline for comparing similar predicates. We show that re-weighting word embeddings and performing PCA for sparse features before fusing them significantly enhances the clustering perfor-

mance, reaching up to 0.58 pairwise F1 score.

## References


Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pages 26–31.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *The twentieth international joint conference on artificial intelligence, IJCAI 2007*, pages 2670—-2676, Hyderabad, India.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 28–36.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26-29, 2011, Banff, Alberta, Canada*, pages 113–120.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 355–366.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Ian T. Jolliffe. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP 2012*, pages 523—-534. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Thahir Mohamed, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2011. Discovering relations between noun categories. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1447–1455.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012a. Discovering and exploring relations on the web. *PVLDB*, 5(12):1982–1985.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012b. PATTY: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 1135–1145.

Harinder Pal and Mausam. 2016. Demonyms and compound relational nouns in nominal open IE. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, pages 35–39.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Joint Human Language Technology Conference/Annual*

*Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '13)*.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2300–2305.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. *CoRR*, abs/1603.01648.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2011. Probabilistic matrix factorization leveraging contexts for unsupervised relation extraction. In *Advances in Knowledge Discovery and Data Mining - 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part I*, pages 87–99.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. 2013. Open Information Extraction with Tree Kernels. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, HLT-NAACL 2013*, pages 868–877, Atlanta, Georgia.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466. Association for Computational Linguistics.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2012. Unsupervised relation discovery with sense disambiguation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 712–720. Association for Computational Linguistics.