End-to-End Relation Extraction Using Distant Supervision from External Semantic Repositories

Truc-Vien T. Nguyen and **Alessandro Moschitti**
Department of Information Engineering and Computer Science
University of Trento
38123 Povo (TN), Italy
{nguyenthi,moschitti}@disi.unitn.it

Abstract

In this paper, we extend distant supervision (DS) based on Wikipedia for Relation Extraction (RE) by considering (i) relations defined in external repositories, e.g. YAGO, and (ii) any subset of Wikipedia documents. We show that training data constituted by sentences containing pairs of named entities in target relations is enough to produce reliable supervision. Our experiments with state-of-the-art relation extraction models, trained on the above data, show a meaningful F1 of 74.29% on a manually annotated test set: this highly improves the state-of-art in RE using DS. Additionally, our end-to-end experiments demonstrated that our extractors can be applied to any general text document.

1 Introduction

Relation Extraction (RE) from text as defined in ACE (Doddington et al., 2004) concerns the extraction of relationships between two entities. This is typically carried out by applying supervised learning, e.g. (Zelenko et al., 2002; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005) by using a hand-labeled corpus. Although, the resulting models are far more accurate than unsupervised approaches, they suffer from the following drawbacks: (i) they require labeled data, which is usually costly to produce; (ii) they are typically domain-dependent as different domains involve different relations; and (iii), even in case the relations do not change, they result biased toward the text feature distributions of the training domain.

The drawbacks above would be alleviated if data from several different domains and relationships were available. A form of weakly supervision, specifically named distant supervision (DS) when applied to Wikipedia, e.g. (Banko et al., 2007; Mintz et al., 2009; Hoffmann et al., 2010) has been recently developed to meet the requirement above. The main idea is to exploit (i) relation repositories, e.g. the *Infobox*, x, of Wikipedia to define a set of relation types $RT(x)$ and (ii) the text in the page associated with x to produce the training sentences, which are supposed to express instances of $RT(x)$.

Previous work has shown that selecting the sentences containing the entities targeted by a given relation is enough accurate (Banko et al., 2007; Mintz et al., 2009) to provide reliable training data. However, only (Hoffmann et al., 2010) used DS to define extractors that are supposed to detect all the relation instances from a given input text. This is a harder test for the applicability of DS but, at the same time, the resulting extractor is very valuable: it can find rare relation instances that might be expressed in only one document. For example, the relation *President(Barrack Obama, United States*) can be extracted from thousands of documents thus there is a large chance of acquiring it. In contrast, *President(Eneko Agirre, SIGLEX)* is probably expressed in very few documents, increasing the complexity for obtaining it.

In this paper, we extend DS by (i) considering relations from semantic repositories different from Wikipedia, i.e. YAGO, and (2) using training instances derived from any Wikipedia document. This allows for (i) potentially obtaining training data

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 277–282,
Portland, Oregon, June 19-24, 2011.

for many more relation types, defined in different sources; (ii) meaningfully enlarging the size of the DS data since the relation examples can be extracted from any Wikipedia document [1].

Additionally, by following previous work, we define state-of-the-art RE models based on kernel methods (KM) applied to syntactic/semantic structures. We use tree and sequence kernels that can exploit structural information and interdependencies among labels. Experiments show that our models are flexible and robust to Web documents as we achieve the interesting F1 of 74.29% on 52 YAGO relations. This is even more appreciable if we approximately compare with the previous result on RE using DS, i.e. 61% (Hoffmann et al., 2010). Although the experiment setting is different from ours, the improvement of about 13 absolute percent points demonstrates the quality of our model.

Finally, we also provide a system for extracting relations from any text. This required the definition of a robust Named Entity Recognizer (NER), which is also trained on weakly supervised Wikipedia data. Consequently, our end-to-end RE system is applicable to any document. This is another major improvement on previous work. The satisfactory RE F1 of 67% for 52 Wikipedia relations suggests that our model is also successfully applicable in real scenarios.

1.1 Related Work

RE generally relates to the extraction of relational facts, or world knowledge from the Web (Yates, 2009). To identify semantic relations using machine learning, three learning settings have been applied, namely supervised methods, e.g. (Zelenko et al., 2002; Culotta and Sorensen, 2004; Kambhatla, 2004), semi supervised methods, e.g. (Brin, 1998; Agichtein and Gravano, 2000), and unsupervised method, e.g. (Hasegawa et al., 2004; Banko et al., 2007). Work on supervised Relation Extraction has mostly employed kernel-based approaches, e.g. (Zelenko et al., 2002; Culotta and Sorensen, 2004; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2005; Bunescu, 2007; Nguyen et al., 2009; Zhang et al., 2006). However, such approaches can be applied to few relation types thus distant supervised learning (Mintz et al., 2009) was introduced to tackle such problem. Another solution proposed in (Riedel et al., 2010) was to adapt models trained in one domain to other text domains.

[1] Previous work assumes the page related to the *Infobox* as the only source for the training data.

Algorithm 2.1: ACQUIRE_LABELED_DATA()

$DS = \emptyset$
$YAGO(R) : Instances\ of\ Relation\ R$
for each $\langle Wikipedia\ article : W \rangle \in Freebase$
do {
  $\mathcal{S} \leftarrow set\ of\ sentences\ from\ W$
  for each $s \in \mathcal{S}$
  do {
    $\mathcal{E} \leftarrow set\ of\ entities\ from\ s$
    for each $E_1 \in \mathcal{E}$ **and** $E_2 \in \mathcal{E}$ **and** $R \in YAGO$
    do {
      if $R(E_1, E_2) \in YAGO(R)$
      then $DS \leftarrow DS \cup \{s, R^+\}$
      else $DS \leftarrow DS \cup \{s, R^-\}$
    }
  }
}
return $(\mathcal{DS})$

2 Resources and Dataset Creation

In this section, we describe the resources for the creation of an annotated dataset based on distant supervision. We use YAGO, a large knowledge base of entities and relations, and Freebase, a collection of Wikipedia articles. Our procedure uses entities and facts from YAGO to provide relation instances. For each pair of entities that appears in some YAGO relation, we retrieve all the sentences of the Freebase documents that contain such entities.

2.1 YAGO

YAGO (Suchanek et al., 2007) is a huge semantic knowledge base derived from WordNet and Wikipedia. It comprises more than 2 million entities (like *persons*, *organizations*, *cities*, etc.) and 20 million facts connecting these entities. These include the taxonomic Is-A hierarchy as well as semantic relations between entities.

We use the YAGO version of *2008-w40-2* with a manually confirmed accuracy of 95% for 99 relations. However, some of them are (a) trivial, e.g. *familyNameOf*; (b) numerical attributes that change over time, e.g. *hasPopulation*; (c) symmetric, e.g. *hasPredecessor*; (d) used only for data management, e.g. *describes* or *foundIn*. Therefore, we removed those irrelevant relations and obtained 1,489,156 instances of 52 relation types to be used with our DS approach.

2.2 Freebase

To access to Wikipedia documents, we used Freebase (March 27, 2010 (Metaweb Technologies, 2010)), which is a dump of the full text of all Wikipedia articles. For our experiments, we used 100,000 articles. Out of them, only 28,074 articles contain at least one relation for a total of 68,429 of relation instances. These connect 744,060 entities, 97,828 dates and 203,981 numerical attributes.

Temporal and Numerical Expression

Wikipedia articles are marked with entities like *Person* or *Organization* but not with dates or numerical attributes. This prevents to extract interesting relations between entities and dates, e.g. *John F. Kennedy was born on May 29, 1917* or between entities and numerical attributes, e.g. *The novel Gone with the wind has 1037 pages*. Thus we designed 18 regular expressions to extract dates and other 25 to extract numerical attributes, which range from integer number to ordinal number, percentage, monetary, speed, height, weight, area, time, and ISBN.

2.3 Distant Supervision and generalization

Distant supervision (DS) for RE is based on the following assumption: (i) a sentence is connected *in some way* to a database of relations and (ii) such sentence contains the pair of entities participating in a target relation; (iii) then it is likely that such sentence expresses the relation. In traditional DS the point (i) is implemented by the *Infobox*, which is connected to the sentences by a proximity relation (same page of the sentence). In our extended DS, we relax (i) by allowing for the use of an external DB of relations such as YAGO and any document of Freebase (a collection of Wikipedia documents). The alignment between YAGO and Freebase is implemented by the Wikipedia page link: for example the link *http://en.wikipedia.org/wiki/James_Cameron* refers to the entity *James_Cameron*.

We use an efficient procedure formally described in Alg. 2.1: for each Wikipedia article in Freebase, we scan all of its NEs. Then, for each pair of entities[2] seen in the sentence, we query YAGO to retrieve the relation instance connecting these entities. Note that a simplified version of our approach is the following: for any YAGO relation instance, scan all the sentences of all Wikipedia articles to test point (ii). Unfortunately, this procedure is impossible in practice due to millions of relation instances in YAGO and millions of Wikipedia articles in Freebase, i.e. an order of magnitude of 10^{14} iterations[3].

3 Distant Supervised Learning with Kernels

We model relation extraction (RE) using state-of-the-art classifiers based on kernel methods. The main idea is that syntactic/semantic structures are used to represent relation instances. We followed the model in (Nguyen et al., 2009) that has shown significant improvement on the state-of-the-art. This combines a syntactic tree kernel and a polynomial kernel over feature extracted from the entities:

$$CK_1 = \alpha \cdot K_P + (1 - \alpha) \cdot TK \qquad (1)$$

where α is a coefficient to give more or less impact to the polynomial kernel, K_P, and TK is the syntactic tree kernel (Collins and Duffy, 2001). The best model combines the advantages of the two parsing paradigms by adding the kernel above with six sequence kernels (described in (Nguyen et al., 2009)).

$$CSK = \alpha \cdot K_P + (1-\alpha) \cdot (TK + \sum_{i=1,..,6} SK_i) \quad (2)$$

Such kernels cannot be applied to Wikipedia documents as the entity category, e.g. *Person* or *Organization*, is in general missing. Thus, we adapted them by simply removing the category label in the nodes of the trees and in the sequences. This data transformation corresponds to different kernels (see (Cristianini and Shawe-Taylor, 2000)).

4 Experiments

We carried out test to demonstrate that our DS approach produces reliable and practically usable relation extractors. For this purpose, we test them on

[2] Our algorithm is robust to the lack of knowledge about the existence of any relation between two entities. If the relation instance is not in YAGO, it is simply assumed as a negative instance even if such relation is present in other DBs.

[3] Assuming 100 sentences for each article.

DS data by also carrying out end-to-end RE evaluation. This requires to experiment with a state-of-the-art Named Entity Recognizer trained on Wikipedia entities.

Class	Precision	Recall	F-measure
bornOnDate	97.99	95.22	96.58
created	92.00	68.56	78.57
dealsWith	92.30	73.47	81.82
directed	85.19	51.11	63.89
hasCapital	93.69	61.54	74.29
isAffiliatedTo	86.32	71.30	78.10
locatedIn	87.85	78.33	82.82
wrote	82.61	42.22	55.88
Overall	91.42	62.57	74.29

Table 1: Performance of 8 out of 52 individual relations with overall F1.

4.1 Experimental setting

We used the DS dataset generated from YAGO and Wikipedia articles, as described in the algorithm (Alg. 2.1). The candidate relations are generated by iterating all pairs of entity mentions in the same sentence. Relation detection is formulated as a multiclass classification problem. The *One vs. Rest* strategy is employed by selecting the instance with largest margin as the final answer. We carried out 5-fold cross-validation with the tree kernel toolkit[4] (Moschitti, 2004; Moschitti, 2008).

4.2 Results on Wikipedia RE

We created a test set by sampling 200 articles from Freebase (these articles are not used for training). An expert annotator, for each sentence, labeled all possible pairs of entities with one of the 52 relations from YAGO, where the entities were already marked. This process resulted in 2,601 relation instances.

Table 1 shows the performance of individual classifiers as well as the overall Micro-average F1 for our adapted CSK: we note that it reaches an F1-score of 74.29%. This can be compared with the Micro-average F1 of CK_1, i.e. 71.21%. The lower result suggests that the combination of dependency and constituent syntactic structures is very important: +3.08 absolute percent points on CK_1, which only uses constituency trees.

[4]http://disi.unitn.it/ moschitt/Tree-Kernel.htm

Class	Precision	Recall	F-measure
Entity Detection	68.84	64.56	66.63
End-to-End RE	82.16	56.57	67.00

Table 2: Entity Detection and End-to-end Relation Extraction.

4.3 End-to-end Relation Extraction

Previous work in RE uses gold entities available in the annotated corpus (i.e. ACE) but in real applications these are not available. Therefore, we perform experiments with automatic entities. For their extraction, we follow the feature design in (Nguyen et al., 2010), using CRF++ [5] with unigram/features and Freebase as learning source. Dates and numerical attributes required a different treatment, so we use the patterns described in Section 2.3. The results reported in Table 2 are rather lower than in standard NE recognition. This is due to the high complexity of predicting the boundaries of thousands of different categories in YAGO.

Our end-to-end RE system can be applied to any text fragment so we could experiment with it and any Wikipedia document. This allowed us to carry out an accurate evaluation. The results are shown in Table 2. We note that, without gold entities, RE from Wikipedia still achieves a satisfactory performance of 67.00% F1.

5 Conclusion

This paper proposes two main contributions to Relation Extraction: (i) a new approach to distant supervision (DS) to create training data using relations defined in different sources, i.e. YAGO, and potentially using any Wikipedia document; and (ii) end-to-end systems applicable both to Wikipedia pages as well as to any natural language text.

The results show:

1. A high F1 of 74.29% on extracting 52 YAGO relations from any Wikipedia document (not only from *Infobox* related pages). This result improves on previous work by 13.29 absolute percent points (approximated comparison). This is a rough approximation since on one hand, (Hoffmann et al., 2010) experimented

[5]http://crfpp.sourceforge.net

with 5,025 relations, which indicate that our results based on 52 relations cannot be compared with it (i.e. our multi-classifier has two orders of magnitude less of categories). On the other hand, the only experiment that can give a realistic measurement is the one on hand-labeled test set (testing on data automatically labelled by DS does not provide a realistic outcome). The size of such test set is comparable with ours, i.e. 100 documents vs. our set of 200 documents. Although, we do not know how many types of relations were involved in the test of (Hoffmann et al., 2010), it is clear that only a small subset of the 5000 relations could have been measured. Also, we have to consider that, in (Hoffmann et al., 2010), only one relation extractor is supposed to be learnt from one article (by using *Infobox*) whereas we can potentially extract several relations even from the same sentence.

2. The importance of using both dependency and constituent structures (+3.08% when adding dependency information to RE based on constituent trees).

3. Our end-to-end system is useful for real applications as it shows a meaningful accuracy, i.e. 67% on 52 relations.

For this reason, we decided to make available the DS dataset, the manually annotated test set and the computational data (tree and sequential structures with labels).

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 85–94.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI*, pages 2670–2676.

Sergey Brin. 1998. Extracting patterns and relations from world wide web. In *Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology*, pages 172–183.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of HLT-EMNLP*, pages 724–731, Vancouver, British Columbia, Canada, October.

Razvan C. Bunescu. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of ACL*.

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems (NIPS'2001)*, pages 625–632.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, United Kingdom.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of ACL*, pages 423–429, Barcelona, Spain, July.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) programtasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840, Barcelona, Spain.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of ACL*, pages 415–422, Barcelona, Spain, July.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of ACL*, pages 286–295, Uppsala, Sweden, July.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *The Companion Volume to the Proceedings of ACL*, pages 178–181, Barcelona, Spain, July.

Metaweb Technologies. 2010. Freebase wikipedia extraction (wex), March.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-AFNLP*, pages 1003–1011, Suntec, Singapore, August.

Alessandro Moschitti. 2004. A study on convolution kernels for shallow statistic parsing. In *Proceedings of ACL*, pages 335–342, Barcelona, Spain, July.

Alessandro Moschitti. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceedings of CIKM*, pages 253–262, New York, NY, USA. ACM.

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of EMNLP*, pages 1378–1387, Singapore, August.

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2010. Kernel-based re-ranking for named-entity extraction. In *Proceedings of COLING*, pages 901–909, China, August.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer Berlin / Heidelberg.

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

Alexander Yates. 2009. Extracting world knowledge from the web. *IEEE Computer*, 42(6):94–97, June.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of EMNLP-ACL*, pages 181–201.

Min Zhang, Jian Su, Danmei Wang, Guodong Zhou, and Chew Lim Tan. 2005. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In *Proceedings of IJCNLP'2005, Lecture Notes in Computer Science (LNCS 3651)*, pages 378–389, Jeju Island, South Korea.

Min Zhang, Jie Zhang, Jian Su, , and Guodong Zhou. 2006. A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of COLING-ACL 2006*, pages 825–832.