Distant Supervision for Relation Extraction using Tree Kernels

Azad Abad and Alessandro Moschitti

DISI, University of Trento, Italy
Qatar Computing Research Institute, Qatar
{abad,moschitti}@disi.unitn.it

Abstract. In this paper we define a simple Relation Extraction system based on SVMs using tree kernels and employing a weakly supervised approach, known as Distant Supervision (DS). Our method uses the simple one-versus-all strategy to handle overlapping relations, i.e., defined on the same pair of entities. The DS data is defined over the New York Times corpus by means of Freebase as an external knowledge base, which indicates the relations of some of the entities of the NYT text. Our experiments show that our simple approach performs well in this domain with respect to the current state of the art.

Keywords: Distant Supervision, Relation Extraction, Support Vector Machines, Tree Kernels

1 Introduction

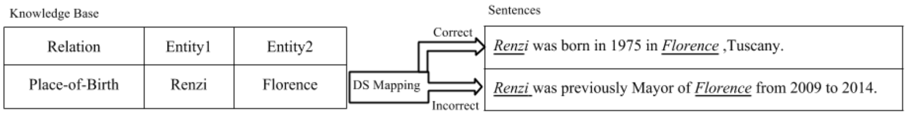
Recently Relation Extraction (RE) has gained popularity in IR community due to its potential applications to question answering, summarization, etc. The RE task concerns extraction of predefined relation types holding between two name entities appearing in text.

In the last decade, supervised learning approaches have been used widely for RE, e.g., [1, 2], obtaining promising results. However, these methods require large-scale human-labeled data to be trained, which is expensive in terms of cost and time. It should be noted that a small dataset can only provide a limited number of relation types, which are not enough to cover application domains.

Distant Supervision (DS) is a new weakly-supervised paradigm designed to overcome the above-mentioned problems [3] by automatically creating labeled training data. This method maps a structured Knowledge Base (KB), e.g., Freebase [1], into a large-scale unlabeled corpus. The mapping is carried out with the following heuristics: given a set of tuples $r{<}e_i{,}e_j{>}\in KB$, if a pair of entities e_i and e_j appear in a text and the same pair participates in an existing relation in KB, the sentence is marked with the corresponding relation label r.

Unfortunately, DS generates noisy training data [11, 6, 5] for the following reasons: (i) the KB can be incomplete and may not contain all the entities that

[1] https://www.freebase.com



Renzi was born in 1975 in *Florence* ,Tuscany.

Renzi was previously Mayor of *Florence* from 2009 to 2014.

Fig. 1. Automatic labeling generated through DS

participate in the relation, thus false negatives can be generated; and (ii) the semantics of the sentence containing the two target entity mentions does not support the relations held by the two entities. For example, let us consider the *Place-of-Birth* relation between <*Renzi*, $Florence$>: the upper sentence of Fig. 1 supports the relation label whereas the lower sentence does not. This generates a training instance misleading the classifier.

In this paper we studied several aspects of DS with the aim of improving its performance, namely we: (i) propose a state-of-the-art machine learning algorithm based on convolution tree kernels for implicitly generating a large amount of structural features, and (ii) demonstrate that, although our method is much easier to re-implement than graphical models [5], it approaches the state of the art.

2 Related Work

So far different approaches have been used for relation extraction and describing all of them is out of the scope of this paper. The most popular approaches are based on supervised learning, where all the instances are manually annotated [1, 2, 8, 7]. The semi-supervised methods use a small number of seed examples manually annotated for extracting patterns from a big corpus [9, 10]. DS was used for the first time to extract binary relation between protein and cells using the Yeast Protein Database (YPD) aligned with scientific articles [3]. Recently, different types of weak labeling resources have been used to address different types of RE problems. In [11] a Multi-Instance Multi-Label (MIML) approach based on undirected graphical model is used to solve the RE problem while in [5] and [13] directed graphical models are proposed. One of the most well-known approaches in supervised and weakly-supervised settings are Tree Kernels (TKs), which have shown promising results in this domain [2, 14–16].

3 SVM and TKs

Support Vector Machine (SVM) is a discriminative classifier that separates classes by providing an optimal hyperplane. One of the valuable features of SVM is the possibility of using kernel functions, which can map instances into high-dimensional feature spaces. Among others, tree kernels have shown their effectiveness in various NLP tasks such as RE [17, 18]. Therefore, we modeled our RE system as a combination of tree kernels applied to syntactic/semantic trees [19] and feature vectors.

3.1 Feature Vector

Our proposed method is based on: 1) tree kernel features and 2) syntactic and lexical features of the two target entity mentions. The latter are originally introduced by the model proposed in [11], which includes: (i) the corresponding Part Of Speech tags (POS), the window of k words of the left and the right of the matched entity mentions, and the sequence of the tokens between them; and (ii) the dependency tree of the sentence. To extract the lexical features, OpenNLP [2] tool was used.

3.2 Tree Kernels

We use the same convolution tree kernels as described in [17] for syntactic parsing. Generally, given two relation examples R_1 and R_2, a composite kernel *K(R_1,R_2)* is computed as:

$$K(R_1, R_2) = \alpha \boldsymbol{x_1} \cdot \boldsymbol{x_2} + (1 - \alpha) K_T(T_1, T_2), \quad (1)$$

$$K_T(T_1, T_2) = \sum_{n_1 \in N1} \sum_{n_2 \in N2} \Delta(n_1, n_2), \quad (2)$$

where α is a coefficient multiplying the target kernel and $\boldsymbol{x_1} \cdot \boldsymbol{x_2}$ is a dot product between two feature vectors of R_1 and R_2. The function $K_T(T_1$,$T_2)$ is a kernel function applied to syntactic trees, where N_1 and N_2 are the set of nodes in the trees T_1 and T_2, respectively and Δ is the number of common sub-trees rooted at n_1 and n_2.

4 Experiment Setup

4.1 Corpus

We considered the New York Times (NYT) corpus provided by [5] to train our system. The corpus originally consists of 1.8 million news articles published between January 1987 and June 2007. Both the training and the test parts are tagged using Stanford Named Entity Recognizer [20], where the training part refers to years 2005-2006 and the test part is extracted from the year 2007. The named entity mentions were mapped with the Freebase knowledge base of 2007 at string level. As a result, 4,700 and 1,900 relations were matched in the training and the test sets respectively. The dataset is highly imbalanced (1:134) in terms of portion of positive and negatives examples. To decrease the effect of dataset disproportion, we only used 50% of examples labeled as NULL relation to train our system.

[2] opennlp.apache.org

Relation Type	P	R	F1
person/nationality	34.5	15.9	21.7
location/contains	10.4	52.2	17.4
person/company	19.8	61.0	29.9
company/place-lived	18.3	10.7	13.5
company-founder	66.7	11.1	19.0

Table 1. Precision and recall of 5 top relation types

	P	R	F1
Mintz++	31.28	15.43	20.67
Surdeanu et al., 2012	29.79	17.48	22.03
Our Model	13.64	29.58	18.68

Table 2. Micro-average for different methods

4.2 Data Processing and models

We used the Charniak parser [21] to generate constituency-based parse trees of the examples. All the name entities were tagged with the Stanford Name Entity Recognizer [20] in four standard classes. We used SVM-Light-TK[3] for learning our models. We applied an SVM binary classifier using the one-vs-all strategy to handle our multi-class classification problem. We tuned the SVM cost-factor parameter (option -j) and the trade-off parameter (option -c) by using 30% of the training set as development set. The effect was an overweighting of the errors on positive examples with respect to errors on negative examples. Indeed, this is crucial when the number of positive examples are dramatically lower than the negative ones (e.g., in NYT corpus some classes have less than 10 examples in the training set). Most importantly, we handle overlapping relations, i.e., when two target entities are in multiple relations, we simply collect the classifier decisions of the one-vs-all method.

4.3 Experimental Evaluation

In order to comparatively evaluate our model, we considered two different state-of-the-art methods: (i) Mintz++ is the baseline of improved version of the original work introduced in [12]. It can predict multiple labels for a given sentence by Or-ing the prediction for the different mentions, (ii) Surdeanu et al. [13] designed a method based on two-level classification, where the first classifies relations whereas the second aggregates them according to a given entity pair.

Tables 1 shows the Precision, Recall and F1 for 5 top classes, which is rather low in line with previous work. More interestingly, Table 2 reports the overall Micro-average Precision, Recall and F1 of Mintz++ and Surdeanu et al. compared with our model. The results show that, although the accuracy of our model is lower than the state of the art, it is promising and it is obtained with a much simpler model and implementation. We point out that our system suffers from low Precision, which depends on the too many false positives caused by the imbalances in the data and also by the high number of incorrect labels assigned by DS to some relation types, when generating training data. This research is under progress and we are going to apply some techniques for decreasing the number of incorrect labels generated by DS.

[3] http://disi.unitn.it/moschitti/Tree-Kernels.htm

5 Conclusion

In this article we presented a simple method for weakly supervised relation extraction. It exploits syntactic information and lexical features by combining tree kernels with feature vectors. Moreover, we apply SVM classifier to handle overlapping relation problem. We experimented with our approach on the well-known RE dataset [5]. The results show that our model can potentially reach the state of the art and at the same time retaining simplicity.

References

1. Culotta, A., Sorensen, J.:Dependency tree kernels for relation extraction.: 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (2004)
2. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction: The Journal of Machine Learning Research, Vol. 3, pp. 1083–1106 (2004)
3. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources: In ISMB, Vol. 1999, pp. 77–86 (1999)
4. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D. S.: Knowledge-based weak supervision for information extraction of overlapping relations.: In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 541–550. Association for Computational Linguistics (2011)
5. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text.: In Machine Learning and Knowledge Discovery in Databases, pp. 148–163 (2010)
6. para la Supervisin, E. M. R., Intxaurrondo, D. A., Surdeanu, M., de Lacalle, O. L., Agirre, E.:Removing Noisy Mentions for Distant Supervision (2013)
7. Kambhatla, N.: Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations.: In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, p. 22. Association for Computational Linguistics (2004)
8. Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R.: Algorithms that learn to extract information.: BBN: TIPSTER phase III, In Proceedings of a workshop on held at Baltimore, 13-15, pp. 75–89. Association for Computational Linguistics, Maryland (1998)
9. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections.: In Proceedings of the fifth ACM conference on Digital libraries, pp. 85–94. ACM (2000)
10. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training.: In Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100. ACM (1998)
11. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D. S.: Knowledge-based weak supervision for information extraction of overlapping relations.: In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 541–550. Association for Computational Linguistics (2011)

12. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data.: In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Vol. 2, pp. 1003–1011. Association for Computational Linguistics (2009)
13. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C. D.: Multi-instance multi-label learning for relation extraction.: In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 455–465. Association for Computational Linguistics (2012)
14. Bunescu, R. C., Mooney, R. J.: A shortest path dependency kernel for relation extraction.: In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 724–73. Association for Computational Linguistics (2005)
15. Nguyen, T. V. T., Moschitti, A.: End-to-end relation extraction using distant supervision from external semantic repositories.: In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, Vol. 2, pp. 277–282. Association for Computational Linguistics (2011)
16. Nguyen, T. V. T., Moschitti, A., Riccardi, G.: Convolution kernels on constituent, dependency and sequential structures for relation extraction.: In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol.3, pp. 1378–1387. Association for Computational Linguistics (2009)
17. Zhang, M., Zhang, J., Su, J., Zhou, G.: A composite kernel to extract relations between entities with both flat and structured features.: In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 825–832. Association for Computational Linguistics (2006)
18. Collins, M., Duffy, N.: Convolution kernels for natural language.: In Advances in neural information processing systems, pp. 625–632 (2001)
19. Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees.: In Machine Learning: ECML, pp. 318–329. Springer, Berlin, Heidelberg (2006)
20. Finkel, J. R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling.: In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. Association for Computational Linguistics (2005)
21. Charniak, E.: A maximum-entropy-inspired parser.: In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pp. 132–139. Association for Computational Linguistics (2000)