From Actors, Politicians, to CEOs: Domain Adaptation of Relational Extractors using a Latent Relational Mapping

Danushka Bollegala
The University of Tokyo
Hongo 7-3-1, Tokyo
113-8656, Japan
danushka@iba.t.u-
tokyo.ac.jp

Yutaka Matsuo
The University of Tokyo
Hongo 7-3-1, Tokyo
113-8656, Japan
matsuo@biz-model.t.u-
tokyo.ac.jp

Mitsuru Ishizuka
The University of Tokyo
Hongo 7-3-1, Tokyo
113-8656, Japan
ishizuka@i.u-tokyo.ac.jp

ABSTRACT

We propose a method to *adapt* an existing relation extraction system to extract new relation types with minimum supervision. Our proposed method comprises two stages: *learning a lower-dimensional projection* between different relations, and *learning a relational classifier* for the target relation type with instance sampling. We evaluate the proposed method using a dataset that contains 2000 instances for 20 different relation types. Our experimental results show that the proposed method achieves a statistically significant macro-average F-score of 62.77. Moreover, the proposed method outperforms numerous baselines and a previously proposed weakly-supervised relation extraction method.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Relation Extraction, Domain Adaptation, Web Mining

1. INTRODUCTION

The World Wide Web contains information related to numerous real-world entities (e.g. persons, locations, organizations, etc.) interconnected by various semantic relations. Accurately detecting the semantic relations that exist between two entities is of paramount importance for information retrieval (IR). For example, to improve coverage in information retrieval, a query about a particular person can return documents describing the various semantic relations that the person under consideration has with other related entities.

Recent work on relation extraction has demonstrated that supervised machine learning algorithms coupled with intelligent feature engineering provide state-of-the-art solutions to this problem [3]. However, supervised learning algorithms depend heavily on the availability of adequate labeled data for the target relation types that must be extracted. Considering the potentially numerous semantic relations that exist among entities on the Web, it is costly to create labeled data manually for each new relation type that we want to extract. Instead of annotating a large set of training data manually for each new relation type, it would be cost effective if

we could somehow *adapt* an existing relation extraction system to those new relation types using a small set of training instances. We study *relation adaptation* – how to adapt an existing relation extraction system that is trained to extract some specific relation types, to extract new relation types in a weakly-supervised setting.

We define **Relation Adaptation** as the problem of learning a classifier for a target relation type $\mathcal{T}$, for which we have a few entity pairs as training instances, given numerous entity pairs for some N source relation types, $\mathcal{S}_1, \ldots, \mathcal{S}_N$. We use the notation $\Omega = \{\mathcal{S}_1, \ldots, \mathcal{S}_N, \mathcal{T}\}$ to denote the set of all relations. A particular relation type from this set is denoted by $\mathcal{R}$ (i.e $\mathcal{R} \in \Omega$). An entity pair that consists of two entities A and B is denoted as (A, B). Moreover, we use the notation $(A, B) \in \mathcal{R}$ to indicate that the relation $\mathcal{R}$ exists between two entities A and B.

2. METHOD

Given a pair of entities (A, B), the first step is to express the relation between A and B using some feature representation. Lexical or syntactic patterns have been successfully used in numerous natural language processing tasks involving relation extraction such as extracting hypernyms [4]. Following the previous work on relation extraction between entities, we use lexical and syntactic patterns extracted from the contexts in which two entities co-occur to represent the semantic relation that exists between those entities. First, we download Web snippets for the AND query of the two entities A and B. Next, we replace A and B respectively by two variables X and Y. Finally, we generate subsequence patterns that contain both X and Y. We extract both lexical and syntactic subsequence patterns using an improved version of the subsequence pattern mining algorithm proposed by Bollegala et al. [2].

We propose a strategy for selecting relation independent patterns using the entropy of a pattern over the distribution of entity pairs. The proposed strategy is inspired by the fact that if a pattern is relation-independent, then its distribution over the entity pairs tends to become more uniform. However, if a pattern is relation-specific, then its distribution is concentrated over a small set of entity pairs that belong to a specific relation type. The entropy of a pattern increases as its distribution becomes more uniform.

We construct a bipartite graph, $G = (V_{RS} \cup V_{RI}, E)$ between relation-specific (V_{RS}) and relation-independent (V_{RI}) patterns to represent the intrinsic relationship between those patterns. Each vertex in V_{RS} corresponds to a relation-specific pattern, and each vertex in V_{RI} corresponds to a relation-independent pattern. A vertex in V_{RS} (corresponding to a relation-specific pattern) is connected to a vertex in V_{RI} (corresponding to a relation-independent pattern) by an undirected edge $e_{ij} \in E$. Note that there are no intra-set edges connecting vertices in V_{RS} and V_{RI}. Moreover,

Table 1: Macro-average results for various methods.

Method	F-measure
Random	7.24
RS patterns	41.41
RI patterns	51.40
All patterns	47.94
Projected	44.86
Combined (all patterns + projected)	56.99
RS patterns + Sampling	49.78
RI patterns + Sampling	54.83
All patterns + Sampling	57.62
Projected + Sampling	47.61
Jiang [5]	55.62
Combined + Sampling (**PROPOSED**)	**62.77**

each edge $e_{ij} \in E$ is associated with a non-negative weight m_{ij}, that measures the strength of association between the corresponding patterns ρ_i and ρ_j. We set m_{ij} to the number of different entity pairs from which both ρ_i and ρ_j are extracted. Edge weights m_{ij} are represented collectively by an edge-weight matrix $\mathbf{M}$ of the bipartite graph G. For simplicity, we use the number of different entity pairs from which two patterns are extracted as the edge-weighting measure. Given as input an edge-weight matrix $\mathbf{M}$ for the bipartite graph G and dimensionality $k(< n)$ of the latent space, we use spectral clustering to compute a projection matrix $\mathbf{U}$ from the original n dimensional pattern space to a k dimensional latent space. The low-dimensional projection reduces the mismatch between patterns in source and target relation types, thereby enabling us to train a classifier for the target relation type using labeled entity pairs for both source and target relation types.

In relation adaptation, the number of target relation training instances (entity pairs) is significantly smaller than that of the source relations. Consequently, most supervised classification algorithms treat the minority class (target relation) instances as noise or outliers. Therefore, learning a classifier for a target relation type which has only a few instances is difficult in practice. To overcome this problem, we use one-sided under-sampling which first selects a subset of the source relation training data and then uses that subset to train a multi-class classifier. One-sided under-sampling methods have been used to select a subset of the majority class in previous work investigating the problem of machine learning with unbalanced datasets [6, 7].

3. EXPERIMENTS AND RESULTS

To evaluate the proposed method, we select 20 relation types that have been used frequently for evaluating relation extraction systems [1, 2] from the YAGO ontology[1]. For each selected relation, we randomly selected 100 entity pairs listed for that relation in the YAGO ontology. Overall, the dataset contains 2000 (20 relations $\times$ 100 instances) entity pairs. The YAGO ontology has a high level of manually confirmed accuracy. It is suitable as a gold standard for evaluating relations between entity pairs on the Web [8].

For each relation type $\mathcal{R}$, we randomly allocated its 100 instances (entity pairs) into three groups: 60 instances as training instances when $\mathcal{R}$ is a source relation, 10 instances as training instances when $\mathcal{R}$ is a target relation, and 30 instances as test instances for $\mathcal{R}$. For each target relation type, therefore we have 1140 (19 $\times$ 60) source relation training instances and 10 target relation training instances, which well simulates the problem setting in relation adaptation. We repeat the above-described data splitting and report the average results of 5 random times.

From Table 1, we see that the proposed method has the best macro-average F-measure among all the different methods. In particular, improvement against the previously proposed state-of-the-art weakly-supervised relation extraction method [5] is statistically significant (paired t-test with $p < 0.05$ inferred as significant). The **Random** baseline on this balanced dataset only yields a very low F-score of 7.25. The **RI patterns** baseline that uses only relation-independent patterns outperforms the **RI patterns** baseline that uses only relation-specific patterns. Using all the patterns (i.e. **All patterns** baseline) performs slightly worse than when using only relation-independent patterns. One reason for this is that the overall performance of the **All patterns** baseline is dominated by the numerous relation-specific patterns, which adapt poorly to target relations. There can be errors in identifying relation-independent patterns using strategies such as mutual information, which engender some noise in the constructed bipartite graph. Consequently, using only the **Projected** features is not satisfactory. However, by augmenting the original features to the projected features (i.e. **Combined** baseline), this problem can be overcome. Next, we evaluate the effect of the one-sided undersampling on top of the numerous baselines discussed above. From Table 1, it is apparent that, by sampling, we consistently improve all the baselines: **RS patterns**, **RI patterns**, **All patterns**, and **Projected**. In fact, the proposed method, which uses augmented feature vectors with sampling, shows a 6 percent improvement over not using sampling (i.e. **Combined** baseline). Moreover, we experimentally verified that the proposed method performs consistently under different parameter settings.

4. CONCLUSION

We proposed and investigated a method to learn a relational classifier for a target relation using multiple source relations. Our experimental results show that the proposed method significantly outperforms 10 baselines and a previously proposed weakly-supervised relation extraction method on a dataset that contains 2000 entity pairs for 20 different relation types. Both feature projection and sampling positively contribute to the proposed method. In future studies, we intend to apply the proposed method to other classification tasks.

5. REFERENCES

[1] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI'07*, pages 2670–2676, 2007.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *WWW'10*, pages 151 – 160, 2010.

[3] Z. GuoDong, S. Jian, Z. Jie, and Z. Min. Exploring various knowledge in relation extraction. In *ACL'05*, pages 427 – 434, 2005.

[4] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING'92*, pages 539–545, 1992.

[5] J. Jiang. Multi-task transfer learning for weakly-supervised relation extraction. In *ACL'09*, pages 1012–1020, 2009.

[6] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML'97*, pages 179 – 186, 1997.

[7] F. Provost. Machine learning from imbalanced data sets. In *AAAI'00 Workshop on Imbalanced Data Sets*, 2000.

[8] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *WWW'07*, 2007.

---

[1] http://www.mpi-inf.mpg.de/yago-naga/yago/