REACTOR: A Framework for Semantic Relation Extraction and Tagging over Enterprise Data

Wei Shen[1], Jianyong Wang[1], Ping Luo[2], Min Wang[2], Conglei Yao[2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]HP Labs China, Beijing, China
[1]chen-wei09@mails.tsinghua.edu.cn, jianyong@tsinghua.edu.cn
[2]{ping.luo, min.wang6, conglei.yao}@hp.com

ABSTRACT

Relation extraction from Web data has attracted a lot of attention in recent years. However, little work has been done when it comes to relation extraction from enterprise data regardless of the urgent needs to such work in real applications (e.g., E-discovery). In this paper, we propose a novel unsupervised hybrid framework, called REACTOR (abbreviated for a fRamework for sEmantic relAtion extraCtion and Tagging Over enteRprise data). We evaluate REACTOR over a real-world enterprise data set and empirical results show the effectiveness of REACTOR.

Categories and Subject Descriptors

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

General Terms

Algorithms,Performance,Experimentation

Keywords

Relation extraction, relation tagging, enterprise data

1. INTRODUCTION

Relation Extraction (RE) is an important research area not only for information retrieval but also for Web mining and question answering. While most work on relation extraction has been focused on Web data, the amount of enterprise data has grown significantly during the past several years for all companies. These enterprise data contain numerous real-world entities and these entities are connected by various types of relations. To make use of such rich information, it is desirable to build an entity relationship graph that can support efficient retrieval of entities and their relations. To achieve this goal, semantic relation extraction from enterprise data is an essential step.

However, the existing techniques on relation extraction cannot be applied directly to enterprise data due to the differences in the data characteristics: the enterprise data has much lower redundancy than Web data. Most existing techniques rely on the high-redundancy nature of the Web data for an ample supply of related entities to achieve reasonable recall. The recall will fall dramatically when applying such techniques to the low-redundancy enterprise data.

Copyright is held by the author/owner(s).
WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

In this paper, we propose a novel unsupervised framework called REACTOR. It uses a statistical method in conjunction with classification and clustering techniques to extract semantic relations and can label the extracted relations with representative tags over enterprise data. Given an enterprise data set, REACTOR first adopts a statistical method to extract a set of representative entity pairs which contain both positive and negative examples for the classifier. Then we extract some features from the positive and negative examples to train the classifier which is in turn used to classify all the other entity pairs each of which appears in the same sentence as related or not. For each entity pair classified as related, a context vector consisting of the words formed from all its occurring sentences is generated, and a clustering algorithm is used to identify the semantic relations of entity pairs. Furthermore, to describe the relations for the entity pairs in each cluster, REACTOR employs a closed frequent sequence pattern mining algorithm to extract the representative tags.

2. THE REACTOR FRAMEWORK

Given a text corpus, the goal of REACTOR is to extract all semantic relations between any two types of entities. We assume that entities of the two corresponding types in this corpus, T_m and T_n, are previously detected like many other relation extraction systems [1]. Figure 1 depicts the architecture of REACTOR, which has four modules including Seed Extractor, Classifier, Cluster, and Relation Tagging.

The Seed Extractor uses statistics to extract a set of representative entity pairs containing both positive and negative examples for the classifier. Specifically, the Seed Extractor computes the relatedness weight for each entity pair $\langle e_i, e_j \rangle$ to assess the probability whether a relationship exists between these two entities as follows:

$$weight(e_i, e_j) = C(e_i, e_j) \log_2 \frac{C(e_i, e_j)}{C(e_i)C(e_j)} \quad (1)$$

where $C(e_i, e_j)$ is the number of co-occurrences of entities e_i and e_j, and $C(e_i)$, $C(e_j)$ are the numbers of occurrences of entity e_i, e_j respectively in the corpus.

Let p_{ij} be an entity pair $\langle e_i, e_j \rangle$ which occurs within one sentence and $P = \{p_{11}, p_{12}, \ldots, p_{ij}, \ldots\}$ be the set of all such entity pairs in the corpus. According to Formula 1, we can calculate the relatedness weight w_{ij} for each entity pair p_{ij} which can tell us how strongly the entity pair p_{ij} is related. Then, the task of extracting positive seeds is to extract a subset of k entity pairs $P' \subseteq P$, such that $\forall p_{ij} \in P'$ and $\forall p_{sv} \notin P'$, we have $w_{ij} \geq w_{sv}$. On the contrary, the task of extracting negative seeds is to extract a subset

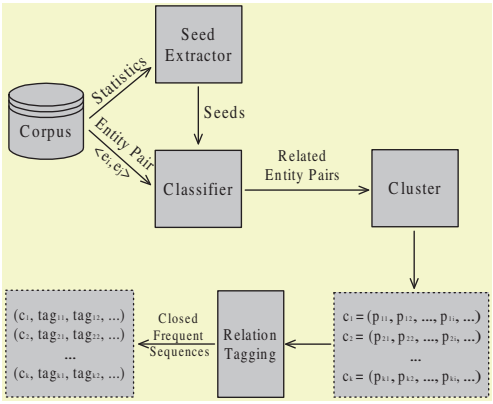

Figure 1: The REACTOR Architecture

of m entity pairs $P'' \subseteq P$, such that $\forall p_{ij} \in P''$ and $\forall p_{sv} \notin P''$, we have $w_{ij} \leq w_{sv}$.

Then we use the seeds produced by Seed Extractor to train the classifier which is *libsvm*[1] used in the experiment. We define some domain-independent features which can be used to capture the syntactic information and entity information for each sentence where the entity pair occurs. The features include: a) The distance between the two entities; b) The type of the entity which appears first in the sentence; c) The part-of-speech tag sequences around the entities; d) The position and type of the other entities in the sentence. Then we use OpenNLP toolkits[2] to annotate each sentence in the entire corpus with POS tags. Finally, each occurrence of the entity pair is presented to the trained classifier and the classifier labels each of them as related or unrelated.

After the classification, we get all related entity pairs. To extract the semantic relations, we assume that entity pairs occurring in the similar context likely have the same semantic relation and can be clustered into a group. We adopt a vector space model to represent the context of an entity pair. Before generating context vectors, we eliminate some non-essential phrases, such as stop words, prepositional phrases and modifiers. Meanwhile, we consider not only the bag of words between the entities but also those around the entities in the sentence within some distance. After generating the context vector for each entity pair, we use cosine similarity to measure the similarity between any two context vectors and then adopt the hierarchical clustering algorithm to further group the entity pairs. Finally, the entity pairs clustered into the same group are expected to have the same semantic relation.

To label the extracted relations, Relation Tagging module employs a closed frequent sequence mining algorithm to identify the closed frequent sequential patterns in all co-occurring sentences where the entity pairs of each cluster appear. In our work, we employ the BIDE algorithm to discover closed sequential patterns. Then we use these extracted patterns to label and describe the semantic relation held in each cluster.

3. EXPERIMENTS

To evaluate the performance of REACTOR, we tested it on a large real-world enterprise data set from HP in which there are over three million pages. We compared REACTOR with the clustering-based method proposed in [1]. The usual metrics of Precision (P), Recall (R) and F-scores (F) on the clustering results were used to evaluate the performance of REACTOR in the same way as that in [1]. In the experiments, we considered the relations in two different domains. One is the PEOPLE-ORGANIZATION (PEO-ORG) domain and another is the PEOPLE-PRODUCT (PEO-PRO) domain. To compare REACTOR with the baseline method and give a quantitative analysis, we randomly selected 500 entity pairs in PEO-ORG domain and 250 entity pairs in PEO-PRO domain respectively as the test data set.

Table 1: Evaluation results on the test data set

Domain	Method	P	R	F
PEO - ORG	Baseline	0.517	0.601	0.556
	Ba-Optimal	0.638	0.549	0.590
	REACTOR	**0.795**	**0.819**	**0.807**
PEO - PRO	Baseline	0.608	0.718	0.659
	Ba-Optimal	0.696	0.670	0.683
	REACTOR	**0.846**	**0.729**	**0.783**

The baseline method needs three parameters including maximum context word length, the occurrence frequency threshold of entity pairs and the norm threshold for context vectors to filter the unreliable pairs. If we use the original setting of these thresholds introduced in [1], all entity pairs in the test data set will be filtered out empirically and no entity pair is retained to start the clustering process, which also strongly reveals the low redundancy of the enterprise data and that the methods based on the high redundancy of Web corpus are not suitable to be applied to enterprise data set. Thereby, to compare REACTOR with the baseline method, we must change the threshold setting of the baseline method. The simplest way is to directly eliminate those thresholds and all entity pairs are retained for the clustering process which we refer to **Baseline**. And we also selected the optimal thresholds for the baseline method which can obtain the best F-scores. This method is referred to **Ba-Optimal**. The optimal thresholds are 10 in PEO-ORG domain and 15 in PEO-PRO domain for the maximum context word length, and 0 for both the occurrence frequency threshold and the norm threshold for both domains. Table 1 shows the experimental results of three different approaches in two different domains. It can be seen from the results that the overall Precision, Recall and F-scores of REACTOR are significantly better than both Baseline and Ba-Optimal in two different relation extraction tasks.

4. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program) under Grant No. 2011CB302206, National Natural Science Foundation of China under grant No. 60833003, an HP Labs Innovation Research Program award, the Okawa Foundation Research Grant, and the Program for New Century Excellent Talents in University under Grant No. NCET-07-0491, State Education Ministry of China.

5. REFERENCES

[1] T. Hasegawa, S. Sekine, and R. Grishman. Discovering relations among named entities from large corpora. In *Proceedings of ACL*, pages 415–422, 2004.

[1]http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
[2]http://opennlp.sourceforge.net/