論文の内容の要旨

Title: Graph Based Learning Methods For Semantic Relation Extraction

（意味的関係抽出のためのグラフに基づく学習手法）

李　海博

For many real world applications, background knowledge is intensively required. The acquisition of relational domain knowledge is still an important problem. Relation extraction systems extract structured relations from unstructured sources such as documents or web pages. These structured relations are as useful as knowledge. Acquiring relational facts like *Acquirer—Acquiree* relation or *Person—Birthplace* relation with a small number of annotated data could have an important impact on these applications such as business analysis research or automatic ontology construction.

The World Wide Web contains a significant amount of relational information expressed in natural language. Although, the Web forms a fertile source of data for relation extraction, the users of relation extraction system are typically required to provide a large amount of annotated texts to identify the interesting relation. This requirement is typically neither feasible nor inapplicable. Therefore, bootstrapping systems are proposed to address the task of Web-based relation extraction, which usually only need a small number of seed entity pairs of relations.

In this thesis, graph based semi-supervised learning methods are applied to improving the performance of bootstrapping relation extraction system. Semi-supervised learning approaches aim at obtaining good performance at a low cost by combining (potentially large) amounts of unlabeled data with labeled data. The bootstrapping relation extraction can be naturally treated as semi-supervised learning problem. The experimental work of this thesis shows that considerable improvements are achieved by using graph based semi-supervised methods.

In Chapter 2, we overview information extraction task and systematically revisit related work on semantic relation extraction problem. There are three kinds of methods to extract relations from documents: traditional relation extraction, open relation extraction and bootstrapping relation extraction. For the traditional relation extraction system, the user is usually required to provide a large amount of annotated texts to identify the interesting relation. On the other hand, the open information extraction system uses some generalized patterns or a small set of relation-independent heuristics to extract all potential of relations between name entities. The bootstrapping relation extraction is a tradeoff between traditional relation extraction and open relation extraction, which uses given seeds to bootstrap relevant instances from the Web or large corpora. The target relations are “weakly” defined by the given seeds.

In Chapter 3, we survey the semi-supervised learning methods used to the information

extraction or text mining task, especially the co-training algorithm and label propagation algorithm. The co-training algorithm and label propagation algorithm are based respectively on different background assumptions. The co-training algorithm is grounded on compatibility assumption. The compatibility assumption means that for any data point x = (x_1, x_2), the classifier f_1 trained on $view_1$ gives x_1 a label that is the same as x_2's, which is given by the classifier f_2 trained on $view_2$. However, one cannot expect that all data are sufficiently compatible in practice. Label propagation is based on the consistency assumption: that nearby points are inclined to achieve the same label. In this thesis, to reduce the influences of incompatibility, the consistency assumption is regarded as a relaxation of the compatibility assumption because a node in one view is merely inclined to, but need not necessarily get, the same label with linked nodes in different views. Based on this idea, we propose a multi-view algorithm in Chapter 5.

Chapter 4 focuses on a bootstrapping relation extraction framework which is mainly composed of two aspects: expanding and ranking. Given the seeds of target relation (e.g. entity pairs or context patterns or both), the expanding component alternately extracts some entity pairs and context patents of target relation. Since some extracted entity pairs and context patterns are irrelevant or weakly relevant to the target relation, it is important for user to put the most relevant instances on the top of a returned list. We propose a ranking module to rank these instances according to their similarities to the given seeds.

In Chapter 5, we propose a graph based multi-view learning algorithm. This algorithm is based on the generalized consistency assumption. This assumption is composed of two parts: intra-view consistency and inter-view consistency. The intra-view consistency means that, in each view, nearby points or points on the same structure are likely to have the same label. The inter-view consistency presume that points from different view co-occurred frequently are inclined to belong to the same class. Different from co-training style algorithm, the proposed algorithm does not require the establishment of compatibility assumption. In most real-world applications, the compatibility assumption is quite strong, since data points from different views might belong to different classes. The co-training algorithm propagates “hard” labels from one view to the other using the compatibility assumption. Instead of propagating the “hard” label on a bipartite graph, the proposed algorithm spreads label scores among different views to avoid view incompatibility.

In Chapter 6, we evaluate the proposed multi-view learning algorithm using the bootstrapping relation extraction framework. We compare the proposed algorithm to the existing methods, relevant score based methods and frequency based methods, the results indicate that the multi-view learning algorithm can improve the performance of the relation extraction systems.

In Chapter 7, the multi-view learning algorithm is applied to semantic relation classification task. It shows that our proposed algorithm can improve label propagation algorithm with single view on the CDL corpus and SemEval-07 dataset. The experimental results also show the robustness of the proposed method to different inter-view correlation measures and different feature splitting.