Towards Large-Scale Unsupervised Relation Extraction from the Web

Bonan Min, New York University, USA
Shuming Shi, Microsoft Research Asia, China
Ralph Grishman, New York University, USA
Chin-Yew Lin, Microsoft Research Asia, China

ABSTRACT

The Web brings an open-ended set of semantic relations. Discovering the significant types is very challenging. Unsupervised algorithms have been developed to extract relations from a corpus without knowing the relation types in advance, but most rely on tagging arguments of predefined types. One recently reported system is able to jointly extract relations and their argument semantic classes, taking a set of relation instances extracted by an open IE (Information Extraction) algorithm as input. However, it cannot handle polysemy of relation phrases and fails to group many similar (“synonymous”) relation instances because of the sparseness of features. In this paper, the authors present a novel unsupervised algorithm that provides a more general treatment of the polysemy and synonymy problems. The algorithm incorporates various knowledge sources which they will show to be very effective for unsupervised relation extraction. Moreover, it explicitly disambiguates polysemous relation phrases and groups synonymous ones. While maintaining approximately the same precision, the algorithm achieves significant improvement on recall compared to the previous method. It is also very efficient. Experiments on a real-world dataset show that it can handle 14.7 million relation instances and extract a very large set of relations from the Web.

Keywords: Information Extraction, Large-Scale, Relation Extraction, Semantics, Unsupervised Learning, Web

INTRODUCTION

Relation extraction aims at discovering semantic relations between entities. It is an important task that has many applications in answering factoid questions, building knowledge bases and improving search engine relevance. In the era of the Internet, the Web has become a massive potential source of such relations. However, there are challenges for Web-scale open-domain relation extraction: the huge and fast-growing scale, a mixed genre of documents and potentially infinite types of relations it carries. To extract these relations, a system should
not assume a fixed set of relation types, nor rely on a fixed set of relation argument types. It also should be able to efficiently handle a very large amount of data.

The past decade has seen some promising solutions. Unsupervised relation extraction (URE) algorithms extract relations from a corpus without knowing the relations in advance. However, most algorithms (Hasegawa, Sekine, & Grishman, 2004; Shinyama & Sekine, 2006; Chen, Ji, Tan, & Niu, 2005) rely on tagging predefined types of entities as relation arguments, and thus are not well-suited for open domain relation extraction.

Recently, Kok and Domingos (2008) proposed Semantic Network Extractor (SNE), which generates argument semantic classes and sets of synonymous relation phrases at the same time. It avoids the requirement of tagging relation arguments of predefined types. However, SNE has 2 limitations: 1) following previous URE algorithms, it only uses features from the set of input relation instances for clustering. Empirically we found that it fails to group many relevant relation instances. These features, such as the surface forms of arguments and lexical sequences in between, are very sparse in practice. In contrast, there exist several well-known corpus-level semantic resources that can be automatically derived from a source corpus and are shown to be useful for generating the key elements of a relation: its 2 argument semantic classes and a set of synonymous phrases. For example, semantic classes can be derived from a source corpus with contextual distributional similarity and web table co-occurrences. The “synonymy” problem for clustering relation instances could potentially be better solved by adding these resources. 2) SNE assumes that each entity or relation phrase belongs to exactly one cluster, thus is not able to effectively handle polysemy of relation phrases. An example of a polysemous phrase is be the currency of as in 2 triples <Euro, be the currency of, Germany> and <authorship, be the currency of, science>. As the target corpus expands from mostly news to the open web, polysemy becomes more important as input covers a wider range of domains. In practice, around 22% (in forthcoming sections) of relation phrases are polysemous. Failure to handle these cases significantly limits its effectiveness.

To move towards a more general treatment of the polysemy and synonymy problems, we present a novel algorithm WEBRE for open-domain large-scale unsupervised relation extraction without predefined relation or argument types (initially presented in Min, Shi, Grishman, & Lin, 2012). The major contributions of this work are:

- WEBRE incorporates a wide range of corpus-level semantic resources for improving relation extraction. The effectiveness of each knowledge source and their combination are studied and compared. To the best of our knowledge, it is the first to combine and compare them for unsupervised relation extraction;
- WEBRE explicitly disambiguates polysemous relation phrases and groups synonymous phrases, thus it fundamentally avoids the limitation of previous methods;
- Experiments on the Clueweb09 dataset (lemurproject.org/clueweb09.php) show that WEBRE is effective and efficient. We present a large-scale evaluation and show that WEBRE can extract a very large set of high-quality relations. Compared to the closest prior work, WEBRE significantly improves recall while maintaining the same level of precision. WEBRE is efficient. To the best of our knowledge, it handles the largest triple set to date (7-fold larger than largest previous effort). Taking 14.7 million triples as input, a complete run with one CPU core takes about a day.

RELATED WORK

Unsupervised relation extraction (URE) algorithms (Hasegawa, Sekine, & Grishman, 2004; Chen, Ji, Tan, & Niu, 2005; Shinyama & Sekine, 2006) collect pairs of co-occurring entities as relation instances, extract features for instances
BCWB: A P300 Brain-Controlled Web Browser
[www.igi-global.com/article/bcwb/176733?camid=4v1a](www.igi-global.com/article/bcwb/176733?camid=4v1a)

Modelling Propagation of Public Opinions on Microblogging Big Data Using Sentiment Analysis and Compartmental Models