Information Processing and Management 000 (2017) 1–21

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Self-training on refined clause patterns for relation extraction

Duc-Thuan Vo*, Ebrahim Bagheri

Laboratory of Systems, Software and Semantics (LS^3), Ryerson University, Toronto, ON, Canada

ARTICLE INFO

Article history:
Received 11 August 2016
Revised 14 February 2017
Accepted 21 February 2017
Available online xxx

Keywords:
Relation extraction
Open information extraction
Self-training algorithm
Syntactic parsing
Dependency parsing

ABSTRACT

Within the context of Information Extraction (IE), relation extraction is oriented towards identifying a variety of relation phrases and their arguments in arbitrary sentences. In this paper, we present a clause-based framework for information extraction in textual documents. Our framework focuses on two important challenges in information extraction: 1) Open Information Extraction and (OIE), and 2) Relation Extraction (RE). In the plethora of research that focus on the use of syntactic and dependency parsing for the purposes of detecting relations, there has been increasing evidence of incoherent and uninformative extractions. The extracted relations may even be erroneous at times and fail to provide a meaningful interpretation. In our work, we use the English clause structure and clause types in an effort to generate propositions that can be deemed as extractable relations. Moreover, we propose refinements to the grammatical structure of syntactic and dependency parsing that help reduce the number of incoherent and uninformative extractions from clauses. In our experiments both in the open information extraction and relation extraction domains, we carefully evaluate our system on various benchmark datasets and compare the performance of our work against existing state-of-the-art information extraction systems. Our work shows improved performance compared to the state-of-the-art techniques.

1. Introduction

Relation Extraction (RE) is one of the important tasks in natural language processing, enabling information extraction and knowledge discovery from text. It aims at organizing relevant segments of unstructured text in relation triples that represent the relationship between two arguments through a relation. As part of an effort to infer more complex relational structures, relation extraction techniques aim to steer the extraction process away from the ambiguous extractions of semantic relations. Representing a particular set of relationships between two or more entities in text can be used for querying and automated reasoning. To infer complex relations, several approaches have been proposed, involving supervised learning (Abacha & Zweigenbaum, 2016; Bunescu & Mooney, 2005; Kambhatla, 2004; Ravichandran & Hovy, 2002; Zhou, Qian, & Fan, 2010), semi-supervised learning (Agichtein & Gravano, 2000; Batista, Martins, & Silva, 2015; Pantel & Pennacchiotti, 2006; Vo & Bagheri, 2015), and unsupervised learning methods (Akbik, Visengeriyeva, Herger, Hemsen, & Loser, 2012; Rosenfeld & Feldman, 2007; Turney, 2008; Yao, Riedel, & McCallum, 2012).

Among the supervised methods, Bunescu and Mooney (2005), Kambhatla (2004), Ravichandran and Hovy (2002), and Zhou et al. (2010) have focused on performing language analysis for semantic relation extraction. A running theme among

* Corresponding author.
E-mail addresses: thuanvd@ryerson.ca, thuanvd@gmail.com (D.-T. Vo), bagheri@ryerson.ca, ebrahim.bagheri@gmai.com (E. Bagheri).

http://dx.doi.org/10.1016/j.ipm.2017.02.009
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

these techniques is the capacity to generate linguistic features based on syntactic, dependency, or shallow semantic structures of the text. Espousing these features, the models are subsequently trained to identify instances of entities that are related through relations. Once the identification process is underway, the extractions are classified based on pre-defined relation types. This is a laborious and time-consuming undertaking on the part of these approaches, involving the analysis of vast quantities of sample data.

Bootstrapping based pattern matching approaches have been employed by various researchers (Agichtein & Gravano, 2000; Brin, 1998; Greenwood & Stevenson, 2006; Pantel & Pennacchiotti, 2006) to extract patterns from seed relations. These approaches exploit the concept of information redundancy and hypothesize that similar relations tend to appear in uniform contexts. The work conducted by Batista et al. (2015) showed that semi-supervised bootstrapping techniques could be used for extracting semantic relations from text by iteratively expanding a set of initial seed relationships. In an effort to find similar relationships, these researchers investigated the effectiveness of bootstrapping for relationship extraction using word embeddings. Their model involves the use of a Named Entity Recognition (NER) module along with weak entity linking by matching entity names with Freebase concepts. In Xu, Uszkoreit, and Li (2007) and Xu, Uszkoreit, Krause, and Hong Li (2010), the authors' goal of extracting relations of various complexities is accomplished through bootstrapping with the ability to automatically learn pattern rules from parsed data. These researchers use dependency trees as the input for pattern extraction and work with trees or sub-trees that contain seed arguments. Despite their eagerness to maintain high accuracy, it is difficult to claim with certainty that the identified patterns are indeed accurate. In lieu of this, there is a probability that faulty seeds could potentially be injected into the bootstrapping process.

The presence of Open Information Extraction (OIE) (Banko, Cafarella, Soderland, Broadhead, & Etzioni, 2007; Etzioni, Fader, Christensen, Soderland, & Mausam, 2011; Fader, Soderland, & Etzioni, 2011; Nebot & Berlanga, 2014; Yahya, Whang, Gupta, & Halevy, 2014; Vo & Bagheri, 2016) offers a more nuanced approach that relies minimally on background knowledge and manually labeled training data. In this respect, various types of relations are taken into consideration without the need to restrict the search for pre-specified semantic relations. Banko et al. (2007), Wu and Weld (2010), and Fader et al. (2011) propose to use shallow syntactic representations of natural language text in the form of verbs or verbal phrases and their arguments. There has also been a more intense interest in approaches that employ robust and efficient dependency parsing for relation extraction (Akbik et al., 2012; Corro & Gemulla, 2013; Garcia & Gamallo, 2011; Mausam, Bart, & Soderland, 2012). Various heuristics are utilized to determine relevant segments of information based on shallow semantic representation or dependency parsing analysis by identifying factors that draw attention to whether two chunks of the original sentence exhibit connection, disconnection, or dependence on one another. Nonetheless, one of the serious drawbacks of techniques that are restricted to shallow syntactic and dependency analysis is detecting relations that display no connection between the verb or verbal phrases in the sentence. Existing state-of-the-art OIE systems the like of ReVerb (Fader et al., 2011) and ClausIE (Corro & Gemulla, 2013) extract relations that are mediated by verbs or verbal phrases based on dependency parsing. Despite key advantages to this approach, the failure to extract all potential relations beyond a pre-defined set of relations including syntactic entities such as nouns and adjectives along with a whole range of verbal structures can be problematic. For instance, consider the following sentence, as shown in Fig. 1, '*Maxus Energy Corp. discovered a new oil field in the southeast Sumatra area of Indonesia.*'. In this sentence, the relation between "*southeast Sumatra area*" and "*Indonesia*" cannot be determined by any type of verbs or verbal phrases through either syntactic or dependency parsing.

To address such limitations, we propose a clause-based framework with refinements to the grammatical structure. We use the English clause structure and clause types in an effort to generate propositions that can be deemed as extractable relations. The framework offers a unique advantage in that it is designed to address some of the more pressing limitations inherent in previous OIE systems through the reformation of the grammatical structure obtained from Syntactic Parsing (SP) and Dependency Parsing (DP). Moreover, an initial seed set generated by multiple high-confidence clause patterns is used for later integration into a bootstrapping process for extracting specified relations. Through the iterative expansion of the original seed set, our work allows for an increasing number of seeds to be identified that can ultimately lead to higher confidence relation extraction patterns. In this paper, our most significant contributions are as follows:

- We demonstrate that a clause-based approach with grammatical structure reformation can be a suitable method for open information extraction to address the following limitations:(1) Identifying relations that previous OIE systems have been oblivious to or overlooked altogether, e.g., the relation between "*southeast Sumatra area*" and "*Indonesia*" in the earlier example and (2) Reducing the number of erroneous relation extractions, e.g., the erroneous identification of '*there*' as a subject of a relation in the following sentence: "*In today's meeting, there were four CEOs*".
- We show that our framework is a suitable method of bootstrapping for relation extraction. It automatically builds an initial seed set based on high confidence clause patterns. Through the iterative expansion of the original seed set, the proposed bootstrapping method allows for an increasing number of seeds to be identified that can ultimately lead to higher confidence relation extraction patterns.

In our work, we empirically show that our framework is highly practical toward building systems for information extraction. We evaluated the approach by carrying out two sets of experiments on textual corpora in the form of 1) Open Information Extraction and 2) Bootstrapping Relation Extraction. The first set of experiments reveals that the approach utilized in our work improves the performance of leading OIE systems such as ClausIE (Corro & Gemulla, 2013), OLLIE (Mausam et al., 2012) and ReVerb (Fader et al., 2011). In the second set of experiments, we apply our proposed method on the standard and

Please cite this article as: D.-T. Vo, E. Bagheri, Self-training on refined clause patterns for relation extraction, Information Processing and Management (2017), http://dx.doi.org/10.1016/j.ipm.2017.02.009