Expert Systems With Applications 60 (2016) 107–117

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Learning to extract domain-specific relations from complex sentences

Saravadee Sae Tan[a,*], Tek Yong Lim[a], Lay-Ki Soon[a], Enya Kong Tang[b]

[a] Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya, 63000, Selangor, Malaysia
[b] Universiti Sains Malaysia, Gelugor, 11800, Pulau Pinang, Malaysia

ARTICLE INFO

Article history:
Received 28 December 2015
Revised 1 May 2016
Accepted 2 May 2016
Available online 3 May 2016

Keywords:
Open information extraction
Structure mapping
Bootstrapping training examples
Greedy mapping

ABSTRACT

Open Information Extraction (OIE) systems focus on identifying and extracting general relations from text. Most OIE systems utilize simple linguistic structure, such as part-of-speech or dependency features, to extract relations and arguments from a sentence. These approaches are simple and fast to implement, but suffer from two main drawbacks: i) they are less effective to handle complex sentences with multiple relations and shared arguments, and ii) they tend to extract overly-specific relations.

This paper proposes an approach to Information Extraction called SemIE, which addresses both drawbacks. SemIE identifies significant relations from domain-specific text by utilizing a semantic structure that describes the domain of discourse. SemIE exploits the predicate-argument structure of a text, which is able to handle complex sentences. The semantics of the arguments are explicitly specified by mapping them to relevant concepts in the semantic structure.

SemIE uses a semi-supervised learning approach to bootstrap training examples that cover all relations expressed in the semantic structure. SemIE inputs pairs of structured documents and uses a Greedy Mapping module to bootstrap a full set of training examples. The training examples are then used to learn the extraction and mapping rules.

We evaluated the performance of SemIE by comparing it with OLLIE, a state-of-the-art OIE system. We tested SemIE and OLLIE on the task of extracting relations from text in the "movie" domain and found that on average, SemIE outperforms OLLIE. Furthermore, we also examined how the performance varies with sentence complexity and sentence length. The results prove the effectiveness of SemIE in handling complex sentences.

1. Introduction and Motivation

The problem of identification and extraction of semantic relations from natural language text has received increasing interest nowadays (Etzioni, Banko, & Cafarella, 2006) (Grishman, 2012). A high-quality, diverse and well-defined knowledge base of extracted information can potentially benefit a wide range of applications. Many knowledge discovery tasks such as information retrieval, question answering, etc may utilize the extracted information to enhance and improve their results. The traditional approaches to Information Extraction (IE), e.g., Snowball (Agichtein & Gravano, 2000), WHISK (Soderland, 1999), rely on hand-crafted rules and hand-tagged training examples to identify and extract specific relations of interest. These approaches focus on a small set of pre-defined relations and require large set of training examples. A recent paradigm, Open Information Extraction (OIE) (Etzioni, Fader, Christensen, Soderland, & Mausam, 2011) (Soderland et al., 2010), facilitates the discovery of domain-independent relations from text and is able to scale to the diversity and size of Web corpus.

Most OIE systems utilize the part-of-speech or dependency features of a sentence to identify the relations and arguments. Noun phrases are classified as arguments and the words between two arguments are utilized as relation phrase. These approaches are simple and fast to implement, but limited to relatively simple sentence structure. Here, we discuss some issues in current state-of-the-art OIE systems. Firstly, OIE systems that use simple linguistic patterns are less effective to handle complex sentences with multiple relations and shared arguments. For example, there are two relations expressed in sentence 1 in Table 1 ("*produced by*" and "*directed by*"). Both relations share the same set of arguments. OLLIE (Mausam, Schmitz, Bart, Soderland, & Etzioni, 2012) is only

* Corresponding author.
E-mail addresses: tansaravadee@gmail.com (S.S. Tan), tylim@mmu.edu.my (T.Y. Lim), lksoon@mmu.edu.my (L.-K. Soon), enyakong1@gmail.com (E.K. Tang).
URL: http://fci.mmu.edu.my/v3/ (S.S. Tan)

http://dx.doi.org/10.1016/j.eswa.2016.05.004
0957-4174/

Table 1
Example problems in OIE.

1	"The film is directed and produced by Wolfgang Petersen."	
	ReVerb:	No extraction
	Ollie:	*be produced by* (*The film; Wolfgang Petersen*) *0.883*
2	"The film is directed by Wolfgang Petersen."	
	ReVerb:	*is directed by* (*The film; Wolfgang Petersen*) *0.977*
	Ollie:	*is directed by* (*The film; Wolfgang Petersen*) *0.907*
3	"Wolfgang Petersen directs the film..."	
	ReVerb:	*directs* (*Wolfgang Petersen; the film*) *0.896*
	Ollie:	*directs* (*Wolfgang Petersen; the film*) *0.798*
4	"Wolfgang Petersen, the director of the film.."	
	ReVerb:	No extraction
	Ollie:	*be the director of* (*Wolfgang Petersen; the film*) *0.623*

able to extract one of the relations and ReVerb (Fader, Soderland, & Etzioni, 2011) extracts none (Table 1). Another issue is that OIE systems tend to extract overly-specific relations. OIE extracts relation triple *rel* (arg_1; arg_2), where *rel* is a phrase describing the relation and arg_1 and arg_2 are arguments. However, two or more relations with different verb phrases may refer to a same concept. In Table 1, sentence 2 and 3 have the same meaning but different syntactic structures. Thus, the relation triples extracted are different. Sentence 4 also has approximately the same meaning but different predicate and thus results in different relation phrase ("*be the director of*"). Furthermore, there is no declaration on the arguments extracted. For sentence 2, the first argument extracted refers to the "*thing directed*" and the second argument refers to the "*director*", whereas sentence 3 is the opposite.

In this paper, we propose a novel and hybrid approach to Open Information Extraction called SemIE (Semantic-based Information Extraction and Mapping). SemIE utilizes a semantic structure, which can be any structure describing the concepts in a domain of discourse. For example, IMDB xml schema (INE, 2011) (Trotman & Wang, 2011) can be used as the semantic structure for the "*movie*" domain. Our approach exploits the predicate-argument structure (PAS) (Kingsbury & Palmer, 2002) (Palmer, Gildea, & Kingsbury, 2005) of a text sentence to identify relations and their arguments, as well as maps the arguments to relevant concepts in the semantic structure. SemIE overcomes the limitation of traditional IE, which heavily relies on hand-tagged training examples for each relation. SemIE uses a semi-supervised learning approach that takes as input pairs of structured documents and uses a Greedy Mapping module to bootstrap a full set of training examples. A pair of structured documents consists of a document annotated with semantic structure and a text annotated with PAS, where both describe a same entity. This approach facilitates an easy collection of training examples that cover all relations expressed in the semantic structure. On the other hand, Open IE tends to extract general relations which are sometimes too generic to provide sufficient information about a domain. SemIE extends OIE by specifying the relevant concepts of the arguments by mapping them to the semantic structure.

The remainder of the paper is organized as follows. In Section 2, we briefly summarize the related work in the area of open information extraction. Section 3 introduces our information extractor, SemIE. We present the experimental results in Section 4. Section 5 concludes with a summary and discussion of future work.

2. Related work

2.1. Open information extraction

Open Information Extraction (OIE) is a new extraction paradigm that facilitates the discovery of domain-independent relations from text and it is able to scale to the diversity and size of the Web corpus. Basically, OIE approaches can be classified according to the level of sophistication of the linguistic structure they rely upon.

2.1.1. Part-of-speech tag

A major category of Open Information Extraction systems such as TextRunner (Banko, Cafarella, Soderland, Broadhead, & Etzion, 2007) (Etzioni, Banko, Soderland, & Weld, 2008), WOE^{pos} (Wu & Weld, 2010) and ReVerb (Fader et al., 2011) makes use of shallow syntactic structure of a text sentence. Most of these approaches aim to extract relation in the form of triple $rel(arg_1; arg_2)$, where arg_1 and arg_2 is a pair of noun-phrase (NP) arguments and *rel* is the relation between them. A classifier is used to predict whether the chosen words between arg_1 and arg_2 indicate a relation or not.

TextRunner (Banko et al., 2007) is one of the first OIE system developed by Etzioni's group. TextRunner uses a self-supervised learning method where training examples are automatically generated based on a set of hand-written rules. Using the training examples, a Naive Bayes classifier is trained based on domain independent features such as POS tag sequences, number of tokens, POS tags of neighboring words, etc. During the extraction process, TextRunner identifies a candidate pair of noun phrase arguments from a sentence based on part-of-speech tagging and noun-phrase chunking. Relation phrase is generated by examining the words in between the arguments. The candidate extraction is then classified as either trustworthy or not.

Wu and Weld (Wu & Weld, 2010) propose WOE (Wikipedia-based Open Extractor) which automatically generates training examples by heuristically matching Wikipedia infobox attributes wih corresponding sentences. WOE^{pos} trains a conditional random field (CRF) classifier based on shallow features like POS tags. The classifier's features include POS tags, regular expressions and conjunctions of features in neighboring words.

StatSnowball (Zhu, Nie, Liu, Zhang, & Wen, 2009) also employs shallow features in their extraction patterns, which are the POS tag sequences between two entities. StatSnowball adopts the bootstrapping approach and applies the ℓ_1-norm regularized maximum likelihood estimation (MLE) to weight the extraction patterns. Starting with a handful set of initial seeds, it iteratively generates new extraction patterns and extracts new relation tuples.

OIE methods such as TextRunner and WOE use "argument first" extraction where a pair of noun phrase arguments are identified first, and words in between the arguments are used to create the relation. Two common types of errors in TextRunner and WOE are incoherent extractions and uninformative extractions. Incoherent extractions are cases where the extracted relation phrase has no meaningful interpretation, whereas uninformative extractions are extractions that omit critical information.

To overcome the problems of incoherent, uninformative and over-specified extractions, ReVerb (Fader et al., 2011) introduces two constraints on binary relations expressed by verbs. The *syntactic constraint* requires a relation phrase to match a POS regular expression. A large dictionary of relations is used as *lexical constraint* to eliminate over-specified relation phrases. ReVerb first identifies relation phrases that satisfy the syntactic and lexical constraints. It then finds a pair of noun-phrase arguments for each identified relation phrase. The resulting extractions are assigned a confidence score using a logistic regression classifier.

ReVerb focuses on identifying a more meaningful and informative relation phrase. However, arguments are generated using simple heuristics such as identifying simple noun phrases on the left and right of a relation phrase. Etzioni et al. add an argument identifier, `ARGLEARNER` to ReVerb (Etzioni et al., 2011). The new system, R2A2, combines ReVerb relation phrase and `ARGLEARNER`'s arguments, which is able to improve the precision and recall values of ReVerb. In Etzioni et al. (2011), a random sample of web