1. Introduction

1.1. Motivation

A large part of the data on the Web is stored in natural language format or unstructured text. While this format provides information targeting towards human consumption, several algorithms for data analysis are not applicable since structured data is required (McCallum, 2005; Cafarella et al., 2011). In order to render a structure from natural language text, the problem of relation extraction stands out as a key problem to find relationships among entities present in the text.

However, the most successful approaches to the relation extraction problem apply supervised machine learning to compute classifiers using features extracted from hand-labeled sentences comprising a training corpus (Nguyen et al., 2007; Zhou et al., 2002; Curran et al., 2003; Finkel et al., 2005). However, supervised methods create several problems, such as a limited number of examples in the training corpus due to an expensive cost of production and domain dependency on corpus annotations. Such limitations prevent this use in constructing Web-scale knowledge bases. An alternative paradigm for relation extraction was introduced by Mintz et al. (2009). The distant supervision approach addresses the problem of creating a considerable number of examples by automatically generating training data from heuristically matching a database relation to text.

On the other hand, the paradigm for publishing data on the Web has been changed from publishing isolated data to publishing data that is linked to other resources and data (Bizer et al. 2009). By doing so, we share knowledge by publishing and accessing documents as part of a global information space. This vision is called the Semantic Web, in which the data published has a structure and

semantics are described by ontologies (Berners-Lee et. al. 2011; Breitman et al. 2006).

In this dissertation, we apply the distant supervision approach along with resources of the Semantic Web to develop a method for relation extraction. We propose a feature based on ontology class hierarchies that, in conjunction with basic lexical features, is used to build a multi-class classifier using supervised machine learning algorithms.

Our experiments evaluate our model using a corpus extracted form the English Wikipedia and instances of the DBpedia Ontology. Experiments are conducted by automatic held-out evaluation and human evaluation. In the held-out experiments, we obtained a total of 88 classes with F-measure greater than 70%. Human evaluation experiments showed that 9 of the 10 top relations, in the number of instances, obtained an accuracy greater than 70%.

1.2. Goal and Contributions

This dissertation demonstrates the importance of Semantic Web resources to natural language processing tasks, using as target the relation extraction problem. We develop a feature based on the hierarchy of classes of instances from a given ontology that are referenced in natural language. We show that the application of such feature can be used to improve the accuracy and recall of classifiers for relation extraction.

1.3. Related Work

The related work comprehends short discussions about supervised methods for relation extraction and addresses their scalability problems. Then, we discuss approaches that propose solutions for scalability such as weak supervision. Finally, we summarize previous applications of resources of the Semantic Web to relation extraction.

Supervised-learning methods were introduced as approaches for information extraction by Soderland et al. (1999) and many other references. Supervised models are the most precise methods for relation extraction (Nguyen et al., 2007; Zhou et al., 2002; Curran et al., 2003; Finkel et al., 2005).

Although supervised learning achieves good precision and recall for the relation extraction task, they are not scalable to thousands of relations on the Web due to the expensive cost of production and the domain dependency on corpus annotations.

In order to tackle the scalability problem on training and testing corpora, the idea of using a database of structured data to heuristically label a textual corpus became popular. It was first introduced as "weak" supervision methods. Craven et al. (1999) applied this strategy to match the Yeast Protein Database to the abstracts of papers in PubMed using a Naive-Bayes based classifier.

Bellare et al. (2007) used BibTex records to train a CRF extractor on 12 bibliographic relations. Wu et al. (2007) used weak supervision to learn relations from the articles using infoboxes of Wikipedia as a database relation. Later, Wu F. et al. (2008) extended this strategy was extended by using smoothing over an automatically generated info box taxonomy.

Mintz et al. (2009) coined the term distant supervision in replacement of the so-called weak supervision. They applied Freebase facts to relational extractors from Wikipedia achieving an average of approximately 67.6% of precision for 100 top relations.

The popularity of distant supervision methods increased rapidly since its introduction. This fact motivated studies regarding the impact of heuristics in the extraction of relation from texts. Unfortunately, depending on the domain of the relation database and text corpus, heuristics can lead to noisy data and poor extraction performance. To alleviate this problem, a distant supervision as a form of multi-instance learning was proposed by Riedel et al (2010). They proposed a new model assuming that at least one of the sentences containing a pair of entities expresses the relation that associates the pair. It is also a contribution of their work that different domains of the database relation and the text corpus increase the

noise generated by heuristics. They showed that the use of Freebase along with a news corpus in distant supervision methods express an average error rate of 18%, more than if used with Wikipedia articles.

A generalization of the method proposed by Riedel et al. (2010) is proposed by Hoffman et al. (2011) in a publicly-available system called MultiR. It implements Riedel et al.'s (2010) method using a probabilistic graphical model that combines a sentence-level extraction component with a corpus-level component for aggregating the individual facts.

Recently, the popularity of the use of resources of the Semantic Web has been increasing. Relation extraction classifiers can be improved by Semantic Web resources as well as new Semantic Web resources can be generated by using relation extraction classifiers. For instance, Gerber et al. (2011) used the DBpedia as a background knowledge to generate several thousands of new facts in DBpedia from Wikipedia articles using distant supervision methods.

1.4. Dissertation Structure

This dissertation is structured as follows. Chapter 2 presents the basic concepts related to this work: the principles, elements and current state of the Semantic Web, Natural Language Processing techniques used as features for our classifier and the theory of Logistic Regression used to train our classifier. Chapter 3 defines the relation extraction problem and describes features applied in our work; it also presents the approach of distant supervision using resources of the Semantic Web and describes the feature space used. Chapter 4 contains experimental results, divided into two classes: held-out evaluation and human evaluation. Finally, Chapter 5 presents the conclusions and the limitations of this work and suggests possible future work.