A Rule-Based Relation Extraction System using DBpedia and Syntactic Parsing

Kamel Nebhi

LATL, Department of linguistics
University of Geneva
Switzerland
kamel.nebhi@unige.ch

Abstract. In this paper, we present a rule-based relation extraction approach which uses DBpedia and linguistic information provided by the syntactic parser Fips. Our goal is twofold: (i) the morpho-syntactic patterns are defined using the syntactic parser Fips to identify relations between named entities (ii) the RDF triples extracted from DBpedia are used to improve RE task by creating gazetteer relations.

Keywords: relation extraction, information extraction, linked open data

1 Introduction

Relation Extraction (RE), defined as the task of recognizing semantic relations between pairs of terms in text, has received renewed interest in the "Web of Data" era, when many billions of RDF triples are actually published on the Linked Open Data (LOD) cloud[1].

While supervised approaches for RE tasks require much human effort, unsupervised approaches need improvement to obtain best results.

In this paper, we propose a rule-based RE method for French exploiting LOD such as DBpedia dataset. Furthermore, the system uses linguistic informations provided by the syntactic parser Fips to define rules.

The main contributions of this paper are twofold: (i) the morpho-syntactic patterns are defined using the syntactic parser Fips to identify relations between named entities (ii) the RDF triples extracted from DBpedia are used to improve RE task by creating a relation gazetteer.

This article is structured as follows: section 2 describes some cognate work on relation extraction; section 3 explains how DBpedia is exploited by our RE approach; section 4 provides details on the proposed approach; section 5 contains our experimental results. We conclude and give some perspectives in section 6.

[1] The interactive view of the Linked Open Data cloud sets is available here : http://lod-cloud.net/

2 Related Work

The researches on RE task are divided on three main approaches: supervised approach, distant supervised approach and unsupervised approach.

Traditional supervised RE has mostly employed kernel-based approaches [20, 14]. But recent work on Open Information Extraction [9, 10] have exploited lexical and syntactic features to solve the problems of incoherent and uninformative extractions. Nevertheless, labeling training data require substantial human effort, leading to significant recent interest in distant supervision.

Distant supervision was introduced in bioinformatics by [6]. Since then, the approach has gained in popularity [4, 12, 11]. The main idea of this approach is to create its own training data by heuristically matching the contents of relation repositories to corresponding text. However, we observe that this method leads to noisy patterns and poor extraction performance when the method is not directly applied to the text we are working with. To solve this problem, [16, 17] use multiple instance learning and multi-instance multi-label learning.

The traditional methods in unsupervised RE collect co-occurrences of word pairs with strings between them, and finally calculate term co-occurrence or generate surface patterns [2, 8]. In addition to surface patterns, [19] generates dependency patterns to obtain semantic information for concept pairs.

3 DBpedia

Linked Open Data refers to data published with a number of best practices based on W3C standards for publishing and connecting structured data on the Web [3]. In the past few years, we have assisted to a growth in the LOD publishing on the Web. Since 2007, many knowledge bases such as DBpedia, Freebase or YAGO have been integrated in the LOD cloud. In this context, exploiting a very large-scale information resource can enhance information extraction process [5].

For our experiments, we used relations extracted from DBpedia French databank[2]. DBpedia [1] defines LOD URIs for millions of concepts by extracting structured information from Wikipedia. The French DBpedia dataset contains data about 100 000 persons, 100 000 locations and 27 020 organizations. Out of them, we extracted 287 976 relation instances.

Table 1 shows details of relations extracted from the French DBpedia dataset.

4 Architecture

Our IE system is built on GATE [7] to annotate entities in text and to detect relations between them. Figure 1 describes the architecture of our RE system.

To start, articles are submitted to an ontology-based Named Entity Recognition (NER) pipeline [13] including a basic linguistic pre-processing, a NE

[2] `http://fr.dbpedia.org/`

Relation Type	Size	Example
per:birthPlace	67 770	Arnold Schwarzenegger, Thal
per:birthDate	64 237	Nelson Mandela, 1918-07-18
per:spouse	7 987	Bill Clinton, Hillary Rodham Clinton
per:residence	3 576	François Hollande, Palais de l'Élysée
location:country	72 468	Wittislingen, Germany
location:mayor	2 237	Chicago, Rahm Emanuel
location:region	51 484	Paris, Île-de-France
org:adminCenter	3 029	UN, New York
org:leaderName	4 080	Thomson Reuters, David Thomson
org:foundedBy	4 016	IBM, Thomas J. Watson
org:foundingYear	4 007	IBM, 1911
org:foundationPlace	3 085	Yahoo, Santa Clara
Total	287 976	

Table 1: The 12 DBpedia relations we use, with their size and an instance of each relation.

gazetteer and rules written in JAPE which is a finite state transducer. Semantic annotation is performed with respect to the DBpedia ontology classes.

The articles are then parallelly processed by the syntactic parser Fips. We chose Fips because it produces syntactic structures with (binary) relations between constituents and because it is robust and accurate enough for our task. NE received from the NER pipeline are submitted to the relation gazetteer. For each sentence and the entity pair on it, the relation gazetteer will identify known relations. Then, we use JAPE patterns based on information, such as functional relations procuding by Fips and/or POS tags, to extract other binary relations.

4.1 The Fips Parser

Fips [18] is a deep symbolic multilingual parser based on generative grammar concepts. The parser uses a bottom up parsing algorithm with parallel treatment of alternatives, as well as heuristics to rank alternatives.

In Fips, each syntactic constituent is represented as a simplified X-bar structure of the form $[_{XP}$ L X R] with no intermediate level, where X is a variable ranging over the set of lexical categories[3]. L and R stand for (possibly empty) lists of, respectively, left and right subconstituents. The lexical level contains detailed morphosyntactic and semantic information available from the manually-built lexicons. In the structures returned by the parser, extraposed elements

[3] The lexical categories are N(oun), Adj(ective), V(erb), P(reposition), Adv(erb), C(onjonction), Inter(jection), to which we add the two functional categories T(ense) and F(unctional).

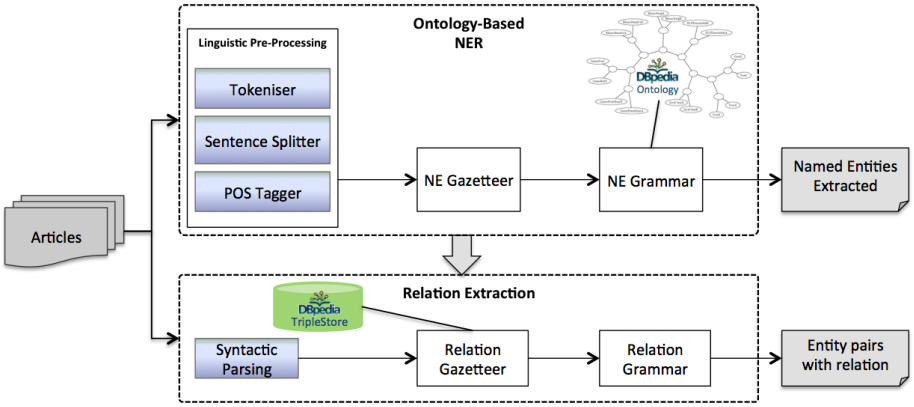

Fig. 1: Relation Extraction Architecture

(interrogative phrases, relative pronouns, clitics, etc.) are coindexed with empty constituents in canonical positions (i.e., typical argument or adjunct positions).

4.2 Lexical and Syntactic Patterns

To characterize binary relationships, we use two different types of relation extraction patterns: lexical patterns, built from words and word class information, and dependency patterns with syntactic information.

Kirkouk	,	dans	le	nord	de	l'	Irak
<city>	,	PRP	DET	nord	PRP	DET	<country>

Table 2: Example Lexical Pattern for the *location:country* relation.

Table 2 shows an example for a simple sentence of a lexical pattern. In other way, Figure 2 provides a dependency pattern of the example (1). The surface strings of the NE have already been replaced by their respectives type in this tree.

(1) *Nestlé a été créée en 1866 à Vevey par Henry Nestlé.*

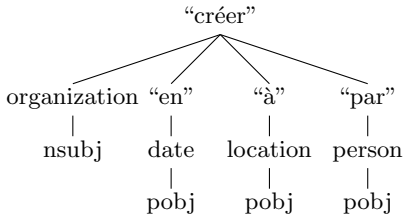

Fig. 2: Dependency Pattern of example (1)

5 Evaluation

The data set we use for our experiments is the Quaero broadcast news extended NE corpus [15]. We have written our rules using a training set of 188 articles. Then, to evaluate the performance of the system we applied the processing resources on a test corpora of 150 articles[4]. We manually constructed relation mentions, according to the DBpedia ontology, only between entity mentions in the same sentence. Then, we compare the system with the gold standard. We present the results according to the different methods in order to examine the contribution of the lexical and syntactic patterns.

RE Model	**Pre%**	**Rec%**	**F1%**
DBpedia	32.2	20.7	25.2
DBpedia+Lex Patterns	49.0	30.0	37.2
DBpedia+Lex&Syn Patterns	75.5	62.1	68.1

Table 3: Results applied to 12 relations for the different RE Model.

Table 3 shows that by using lexical and syntactic patterns, we obtain significant F-measure improvements. The "DBpedia+Lex&Syn patterns" system achieved a F-Measure of 68.1 %.

6 Conclusion - Further Work

In this paper, we propose a novel approach to RE by exploiting syntactic information and DBpedia dataset. We show that this approach provides several advantages and improves RE performance. It was designed mainly to enrich the annotation produced by an ontology-based information extraction system, but can be used for other domains, such as improvement of DBpedia.

[4] For the evaluation, we use relations between Person, Organization, Location and Date.

In future work, we plan to incorporate an anaphora resolution system to improve RE task. We also try to apply the method for other kind of relations.

References

1. Bizer, C. et al.: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the WWW, 2009.
2. Banko, M. et al.: Open information extraction from the Web. In Proceedings of IJCAI, 2007.
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5:1–22, 2009.
4. Bunescu, R., Mooney, R.: Learning to extract relations from the web using minimal supervision. In Proceedings of ACL, 2007.
5. Ciravegna, F., Gentile, A. L., Zhang, Z.: LODIE: Linked Open Data for Web-scale Information Extraction. SWAIE: 11-22, 2012.
6. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In Proceedings of ICISMB, 1999.
7. Cunningham, H. et al.: Text Processing with GATE (Version 6). University of Sheffield, 2011.
8. Davidov, D., Rappoport, A.: Classification of Semantic Relationships between Nominals Using Pattern Clusters. In Proceedings of ACL, 2008.
9. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In Proceedings of EMNLP, 2011.
10. Mausam, Schmitz, M. et al.: Open language learning for information extraction. In Proceedings of EMNLP-CoNLL, 2012.
11. Min, B. et al.: Distant Supervision for Relation Extraction with an Incomplete Knowledge Base, In Proceedings of NAACL-HLT, 2013
12. Mintz, M. et al.: Distant supervision for relation extraction without labeled data. In Proceedings of ACL-IJCNLP, 2009.
13. Nebhi, K.: Ontology-Based Information Extraction for French Newspaper Articles. In Proceedings of KI, 2012.
14. Nguyen, T., Moschitti, A., Riccardi, G.: Convolution kernels on constituent, dependency and sequential structures for relation extraction. In Proceedings of EMNLP, 2009.
15. Rosset, S. et al.: Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers. In Proceedings of LAW VI, 2012.
16. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In Proceedings of ECML, 2010.
17. Surdeanu, M. et al.: Multi-instance Multi-label Learning for Relation Extraction. In Proceedings of EMNLP-CoNLL, 2012.
18. Wehrli, E.: Fips, a deep linguistic multilingual parser: In Proceedings of ACL 2007 Workshop on deep Linguistic Processing, 2007.
19. Yan, Y. et al.: Unsupervised relation extraction by mining Wikipedia texts using information from the web. In Proceedings of ACL and AFNLP, 2009.
20. Zelenko, D., Aone, C., Richardella, A.: Kernel Methods for Relation Extraction. Journal of Machine Learning Research, (3):1083-1106, 2003.