# An Overview of Open Information Extraction

**Pablo Gamallo**

**CITIUS**
**Universidade de Santiago de Compostela**
**Galiza, Spain**
pablo.gamallo@usc.es

**Abstract**

Open Information Extraction (OIE) is a recent unsupervised strategy to extract great amounts of basic propositions (verb-based triples) from massive text corpora which scales to Web-size document collections. We will introduce the main properties of this extraction method.

**1998 ACM Subject Classification** I.2.6 Learning: Knowledge acquisition

**Keywords and phrases** information extraction, natural language processing

**Digital Object Identifier** 10.4230/OASIcs.SLATE.2014.13

**Category** Invited Talk

## 1 Introduction

Recent advanced techniques in Information Extraction aim to capture shallow semantic representations of large amounts of natural language text. Shallow semantic representations are conceived as an intermediate level in the process of structuring textual information. In further processes, shallow semantics can be applied to more complex semantic tasks involved in text understanding, such as textual entailment, filling knowledge gaps in text, or integration of text information into background knowledge bases.

There is an emerging field in Information Extraction interested in applying shallow semantics techniques, namely Machine Reading [7], Learning by Reading [4], or Discovery Information[1]. In this new field, the different techniques used to perform the extraction are not bound by a pre-specified schema of information, but rather they discover relational or categorial structure automatically from given unstructured data using unsupervised strategies.

One of the most used strategies in this new field aimed at discovering shallow semantic representations is known as Open Information Extraction (OIE). The main goal of OIE is to extract a large set of verb-based *triples* (or *propositions*) from unrestricted text. An OIE system reads in sentences and rapidly extracts one or more textual assertions, consisting in a verb relation and two arguments, which try to capture the main relationships in each sentence [3]. Wu and Weld [13] define an OIE system as a function from a document $d$, to a set of triples, $(arg1, rel, arg2)$, where $arg1$ and $arg2$ are verb arguments and $rel$ is a textual fragment (containing a verb) denoting a semantic relation between the two verb arguments. Unlike other relation extraction methods focused on a predefined set of target relations, the Open Information Extraction paradigm is not limited to a small set of target relations known in advance, but extracts all types of (verbal) binary relations found in the text. The main general properties of OIE systems are the following: (i) they are domain independent,

[1] http://www.aha-workshop.de/

3rd Symposium on Languages, Applications and Technologies (SLATE'14).
Editors: Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões; pp. 13–16

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

(ii) they rely on unsupervised extraction methods, and (iii) they are scalable to large amouts of text [6].

## 2 Basic Propositions

An OIE system extracts different triples *(arg1, rel, arg2)*, representing basic propositions or assertions from each sentence in a text. Propositions are defined as coherent and non over-specified pieces of basic information. Consider for example the sentence:

> *In May 2010, the principal opposition parties boycotted the polls after accusations of vote-rigging.*

An OIE system must transform this sentence into a set of triples:

(“*the principal opposition parties*”, “*boycotted*”, “*the polls*”)
(“*the principal opposition parties*”, “*boycotted the polls in*”, “*May 2010*”)
(“*the principal opposition parties*”, “*boycotted the polls after*”, “*accusations of vote-rigging*”)

They represent coherent and non over-specified items of information organized by means of three different relations: “boycotted”, “boycotted the polls in”, and “boycotted the polls after”. Incoherent extractions would be for, instance, the following triples:

(“*parties boycotted*”, “*after*”, “*accusations of vote-rigging*”)
(“*May 2010*”, “*boycotted*”, “*the polls*”)

They are incoherent because, on the one hand, “parties boycotted” cannot be considered as the argument of any relation and, on the other hand, “May 2010” should not be taken as the subject argument of “boycotted”.

Examples of over-specified triples extracted from the same sentences are:

(“*the principal opposition parties*”, “*boycotted the polls in May 2010 after accusations of*”, “*vote-rigging*”)
(“*the principal opposition parties*”, “*boycotted*”,
“*the polls in May 2010 after accusations of vote-rigging*”)

They are over-specified since both the relation “boycotted the polls in May 2010 after accusations of” and the argument “the polls in May 2010 after accusations of vote-rigging” convey too much information to be useful in further semantic tasks, such as semantic entailment or ontology population.

## 3 Overview of Different OIE Systems

A great variety of OIE systems has been developed in recent years. They can be organized in two broad categories: those systems requiring automatically generated training data to learn a classifier and those based on hand-crafted rules or heuristics. In addition, each system category can also be divided in two subtypes: those systems making use of shallow syntactic analysis (PoS tagging and/or chunking), and those based on dependency parsing. In sum, we identify four categories of OIE systems:

**(1) Training data and shallow syntax:** The first OIE system, TextRunner [2], belongs to this category. Two more recent versions of TextRunner, also using training data and shallow syntactic analysis, are ReVerb [9] and R2A2 [8]. Another system of this category is WOE$^{pos}$ [13] whose classifier was trained with corpus obtained automatically from Wikipedia.

**(2) Training data and dependency parsing:** These systems make use of training data represented by means of dependency trees: WOE$^{dep}$ [13] and OLLIE [12].

**(3) Rule-based and shallow syntax:** They rely on lexical-syntactic patterns hand-crafted from PoS tagged text: ExtrHech [15] and LSOE [14].

**(4) Rule-based and dependency parsing:** They make use of hand-crafted heuristics operating on dependency parses: ClauseIE [6], CSD-IE [5], KrakeN [1], and DepOE [11].

## 4 Evaluation and Conclusions

According to the experiments and evaluation we have performed [10], we showed that the rule-based systems perform better than the classifiers based on automatically generated training data. This is in accordance with previous work reported in [6, 5]. Moreover, the systems based on dependency analysis improve over those relying on shallow syntax (TextRunner and ReVerb). It follows that it is not necessary to make use of training data and machine learning strategies to perform open information extraction. We just require a dependency parser and a set of basic rules transforming the parses into triples.

**Acknowledgments.** This work was partially supported by Projects Celtic, Plastic (Innterconecta, FDTI), and HPCPLN (Xunta de Galicia).

## References

1 Alan Akbik and Alexandre Loser. Kraken: N-ary facts in open information extraction. In *Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 52–56, 2012.

2 Michele Banko, , and Oren Etzioni. The tradeoffs between open and traditional relation extraction. In *Annual Meeting of the Association for Computational Linguistics*, 2008.

3 Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *International Joint Conference on Artificial Intelligence*, 2007.

4 K. Barker, B. Agashe, S. Chaw, J. Fan, N. Friedland, M. Glass, J. Hobbs, E. Hovy, D. Israel, D.S. Kim, et al. Learning by reading: A prototype system, performance baseline and lessons learned. In *Proceeding of Twenty-Second National Conference of Artificial Intelligence (AAAI 2007)*, 2007.

5 Hannah Bast and Elmar Haussmann. Open information extraction via contextual sentence decomposition. In *ICSC 2013*, pages 154–159, 2013.

6 Luciano Del Corro and Rainer Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the World Wide Web Conference (WWW-2013)*, pages 355–366, Rio de Janeiro, Brazil, 2013.

7 Oren Etzioni, Michele Banko, and Michael J. Cafarella. Machine reading. In *AAAI Conference on Artificial Intelligence*, 2006.

8 Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open information extraction: the second generation. In *International Joint Conference on Artificial Intelligence*, 2011.

9 Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Conference on Empirical Methods in Natural Language Processing*, 2011.

**10** Pablo Gamallo and Marcos Garcia. Open information extraction based on argument structure detection. In *(submitted paper)*, 2014.

**11** Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. Dependency-based open information extraction. In *ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France, 2012.

**12** Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, 2012.

**13** Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In *Annual Meeting of the Association for Computational Linguistics*, 2010.

**14** Clarissa C. Xavier, Marlo Souza, and Vera S. de Lima. Open information extraction based on lexical-syntactic patterns. In *Brazilian Conference on Intelligent Systems*, pages 189–194, 2013.

**15** Alisa Zhilla and Alexander Gelbukh. Comparison of open information extraction for engish and spanish. In *Dialogue 2014*, 2014.