Xavier *et al. Journal of the Brazilian Computer Society* (2015) 21:4
DOI 10.1186/s13173-015-0023-2

Journal of
the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

Open information extraction based on lexical semantics

Clarissa Castellã Xavier[1*], Vera Lúcia Strube de Lima[1] and Marlo Souza[2]

Abstract

Background: Open Information Extraction (Open IE) aims to obtain not predefined, domain-independent relations from text. This article introduces the Open IE research field, thoroughly discussing the main ideas and systems in the area as well as its main challenges and open issues. The paper describes an open extractor elaborated from the belief that it is not necessary to have an enormous list of patterns or several types of linguistic labels to better perform Open IE. The extractor is based on generic patterns that identify relations not previously specified, including rules corresponding to Cimiano and Wenderoth proposal to learn Qualia structure.

Methods: Named LSOE (Lexical-Syntactic pattern-based Open Extractor) and designed to validate such strategy, this extractor is presented and its performance is compared with two Open IE systems.

Results: The results demonstrate that LSOE extracts relations that are not learned by other extractors and achieves compatible precision.

Conclusions: The work reported here contributes with a new Open IE approach based on pattern matching, demonstrating the feasibility of an extractor based on simple lexical-syntactic patterns.

Keywords: Natural language processing; Information extraction; Open information extraction; Relation extraction

Background

Books and other text documents keep much of the human knowledge. For that reason, it is important to develop computational tools that extract and synthesize information from natural language text with the aim of building large-scale knowledge bases. The task of machine understanding of textual documents mainly parses and transforms unstructured text into a structured representation. This representation should be unambiguous - making it suitable for machine reading and machine interpretation [1].

With the advent of the Semantic Web, the need for methods that perform automatic extraction of semantic data from texts becomes even more relevant [2]. Angeli and Manning [3], for example, propose to use the relations extracted from texts to enlarge databases of known facts and to predict facts, introducing the notion of fact similarity. Novel smartphone interfaces using mining techniques such RevMiner [4] that navigates and analyzes reviews are also based on relations extracted from texts.

Information extraction (IE) systems aim to identify structured relations, like tuples, from unstructured sources such as documents or web pages. IE methods can be used to help building knowledge representation models that report relations between words, like ontologies, semantic networks, and thesauri, among others. According to Fader et al. [5] 'typically, IE systems learn an extractor for each target relation from labeled training examples'. They are usually domain dependent, and their adaptation to a new domain requires manual labor comprising specification and implementation of new patterns of relationships or corpora annotation [6]. Moreover, this approach is not scalable to corpora with a large number of target relationships or where the target relationships cannot be specified in advance [7].

Aiming to overcome this knowledge acquisition bottleneck, the Open IE approach was introduced in 2007 in conjunction with the TextRunner system [8]. According to Li et al. [9] 'Open IE is a domain-independent extraction

*Correspondence: clarissacastella@gmail.com
[1] Faculty of Informatics, Pontifical Catholic University of Rio Grande do Sul, Av. Ipiranga 6681, Porto Alegre, Brazil
Full list of author information is available at the end of the article

paradigm that uses some generalized patterns to extract all the potential relationships between entities'. Wu and Weld [10] define an Open Information Extractor as a function from a document, *d*, to a set of triples in form of (*arg1, rel, arg2*), where *arg1* and *arg2* are noun phrases and *rel* is a textual fragment indicating an implicit, semantic relation between these two noun phrases. It should be noted that in contrast with Traditional Relation Extraction that uses specific types of relations as synonymy, antonymy, hyponymy and lexical inheritance, meronymy, entailment, and presupposition, Open IE considers that all connections among concepts, entities, events, and also those expressed by means of attributes can be considered as relations.

This work focuses on the Open IE paradigm, aiming to highlight its potential and relevance. For that, we present the main studies in this research field and a discussion about evaluation in this area. We bring considerations about the open problems in the area, the main challenges, and open issues. To illustrate this panel, we present our proposal of open extractor, elaborated from the idea that it is not necessary to have an enormous list of patterns or several types of linguistic labels to better perform Open IE. Founded on the work of Pustejovsky [11] on Lexical Semantics and Computational Linguistics and on the learning of Qualia structure from sentences, as proposed by Cimiano and Wenderoth [12], we explore the intrinsic semantic relations between nominals to extract relation tuples from unstructured texts, based on lexical-syntactic patterns tailored to identify such structures.

To validate this proposal, we developed a method to extract relations from POS-tagged texts, using lexical-syntactic patterns. The strategy is constructed on two kinds of patterns: (1) generic patterns to identify domain-specific non-specified relations proposed in the context of our research and (2) rules from Cimiano and Wenderoth proposal [12] to learn Qualia structure which are grounded on Pustejovsky's work [11]. To test this approach, we developed a prototype called LSOE (Lexical-Syntactic pattern-based Open Extractor). LSOE performance was compared to two Open IE systems: ReVerb and DepOE. Our extractor achieved precision compatible to those systems.

The remainder of this paper is organized as follows. Section 'Semantic relations, information extraction, and open information extraction' discusses the methods and concepts related to IE and Open IE, providing the major conceptual divergences between the two paradigms. Section 'Generative lexicon and qualia structure' briefly presents Pustejovsky's generative lexicon theory [11], which contributes to the background of our approach, and the proposal of Cimiano and Wenderoth [12]. Section 'An open extractor based on lexical-syntactic patterns' discusses related work depicting Open IE systems and semantic relation extraction. In Section 'Methods,' we describe our approach and we put on view the lexical-syntactic patterns used to extract relations from text. We also present the prototype developed to evaluate our proposal. In Section 'Results and discussion,' we compare the prototype performance with state-of-the-art Open IE systems and present a discussion on problems and possible solutions. Section 'Conclusions' closes the paper with conclusions and points to future work.

Semantic relations, information extraction, and open information extraction

In this section, we discuss the notion of relation used in Open IE and position it within the grounded literature in the area of automatic relation extraction from texts. Then, we study the conceptual divergences between traditional IE and Open IE. We finish this section presenting Open IE's most relevant works.

Semantic relations in the open ie context

According to Pustejovsky [11], lexical semantics is 'the study of how words are semantically related to one another,' so synonymy, antonymy, hyponymy and lexical inheritance, meronymy, entailment, and presupposition would be examples of interlexical semantic relations. However, the idea of semantic relation in the context of relation extraction goes beyond interlexical semantic relations. As pointed by Nastase et al. [13] 'every nontrivial text describes interactions and relations'. Relations are the connections we perceive among concepts, entities, events, and also those expressed by means of attributes. For example, the sentence *Joe bought a beautiful home* informs relations such as *(Joe, bought, a beautiful home), (Joe, bought, a home), and (beautiful, is a property of, home)*. So, it would be better to use Khoo and Na's concept that states that semantic relations are 'meaningful associations between two or more concepts, entities, or sets of entities' [14].

Thus, what we know about the world consists, in large part, of semantic relations. 'For an automatic system to grasp a text's semantic content, it must be able to recognize and reason the relationships in texts' [13]. Currently, due to the availability of large corpora of texts, mining relations in these texts become more and more frequent. The relational knowledge sought in this case has been of different types, mainly taxonomic knowledge, ontological knowledge, or event knowledge.

Such knowledge contributes to the understanding of relations that occurs in texts, and those relations can in turn become part of the knowledge we hold. According to Murphy [15], semantic relations have the property of uncountability, meaning that there is no objective way to decide the number of relation types, so relations become an open class. This statement has been adopted

by researchers who focus on the extraction of relations from large amounts of texts, including verbal relations, so that the resulting set of relations is an open-ended set. This is the case for the Open IE paradigm and for the present work.

IE and open IE

Information extraction systems aim to identify structured information from unstructured sources such as documents or web pages. More specifically, according to Banko and Etzioni [16], IE 'is the task of recognizing the assertion of a particular relationship between two or more entities in text.'

Work on information extraction from text dates back to the late 70s, implemented by Gerald deJong ([17,18] *apud* [19]). Those first attempts relied on predefined templates and heuristics to provide information extraction. IE has become an important subject within the NLP (natural language processing) community since it was included as a challenge by the MUC conferences in the 80s. Templates and heuristic methods have been applied with relative success over texts in the biomedical or chemistry areas, among others, but proved very difficult to generalize or to adapt in order to evolve to a domain independent approach. More recent attempts handle machine learning methods, hard-coded rules, or a combination of these.

IE deals with the discovery of instances of a prerefined set of relations from a domain. A relation, for the IE literature, is usually described as a tuple $t = (\mathrm{e1}, \ldots, \mathrm{e_n})$ of n entities implicitly or explicitly mentioned in a document or collection, possibly associated to a tag r or a name describing this relation.

The Open IE paradigm was introduced by Banko et al. [8] aiming to develop non-lexicalized, domain-independent extractors of information. The main goal of their work is to provide ways to extract relational information from text in a self-supervised way overcoming the problems of traditional IE methods for scalability and portability across domains. These concerns are confirmed in [20] where Banko identifies the challenges that Open IE must address to perform extraction over very large corpora, namely, automation, domain-independence, and scalability.

The notion of relation under the Open IE paradigm is broader than the one used in IE, as it accounts not just for a tuple of entities, such as *(Aristotle, was born, Stageira)*, what Banko et al. call a concrete tuple [8], but also for a more general kind of relation, defined by those authors as 'unspecified or implying properties of general classes,' as in *(Philosopher, is author of, book)*. We argue that such abstract tuples may be interpreted as semantic relations between nominals, i.e., relations between the concepts associated with the nominals of a given domain, what is also the idea in [10].

According to Banko [20], 'traditional IE methods learn distinct models for individual relations using patterns or labeled instances, requiring manual labor that is linear in the number of relations. Open IE learns a single domain-independent extraction model that discovers an unbounded set of relations with only a one-time cost.' For instance, traditional IE uses as input-labeled data while Open IE uses domain-independent knowledge. The relations learned by traditional IE systems need to be specified in advance, while Open IE systems automatically discover them.

Open IE systems consider that every phrase between a pair of entities can denote a relation. This vision addresses the coverage limitation seen in traditional IE. However, it introduces a substantial amount of noise. In that way, open extractors improve precision by restricting relations to specific part-of-speech tag sequences that are intended to express true relations [1].

In short terms, Open IE algorithms extract relation instances (in the form of triples) from open domain [21]. Taking the relation instances extracted by Open IE algorithms as input, other algorithms have been proposed to resolve objects and relation synonyms [22], extract semantic networks [23], map extracted relations into an existing ontology [24], enlarge databases of known facts, and predict facts [3].

Open IE systems

Unlike traditional IE, which focuses on a predefined set of target relations, Open IE is supposed to extract all kinds of n-ary relations in the text. The goal of an Open IE system is to obtain the largest number of correct triples *(arg1, rel, arg2)* for any relation in the text, where *arg1* and *arg2* are the arguments of the relation and *rel* is a relation phrase.

Open IE systems use three main routes to implement relation extraction. The first one is machine learning, to automatically learn the patterns from a training corpus. The second one is based on heuristics and aims at identifying the occurrence of specific patterns in the text. The last one is the combination of the first two approaches in a hybrid one.

TextRunner [8] introduced the Open IE paradigm. This system uses the machine learning strategy. It accepts POS tagged and noun phrase (NP) chunked sentences as input and analyzes the text between noun phrases to learn relations. For each pair of noun phrases, it applies a CRF classifier[a] to determine if there is a relationship between them. This system was evaluated using a test corpus of 9 million Web documents, obtaining 7.8 million tuples. Human reviewers evaluated a set of 400 randomly selected tuples from which 80.4% were considered correct.

WOE [10] is a continuation of TextRunner including changes in the training data. This new system uses heuristic correspondences between values of attributes

in Wikipedia infoboxes and sentences in order to build training data. It operates in two modes: the first (WOEpos) restricted to POS tagging functions and the second (WOEparse) using dependency parse functions. The paper reports experiments comparing TextRunner, WOEparse, and WOEpos performance. The authors argue that WOEpos reached F-measure between 15% and 34% higher than TextRunner, and that WOEparse reached F-measure between 79% and 90% greater than TextRunner.

Fader et al. [5] and Etzione et al. [7] describe ReVerb system, the first system that uses the heuristic strategy, starting an Open IE second generation of extractors. The system design is based on simple heuristics that identify verbs expressing relationships in English. It receives as input POS tagged and chunked sentences. First, the algorithm identifies the relations and then obtains their arguments. Since the method achieves high recall but low precision, they establish a threshold to assign a confidence score to each extraction. In [7], those authors report that ReVerb achieved an AUC (area under precision-recall curve) twice as big as TextRunner and WOEpos AUC and 38% higher than WOEparse AUC.

Gamallo et al. [25] propose the extraction of relations in other languages than English for the improvement of Open IE methods. They present a system, named DepOE, that uses the heuristic strategy, performing unsupervised extraction of triples using a rule-based analyzer. The extraction method consists of three steps: dependency parsing, identification of clause constituents, and application of extraction rules. Regarding the system performance, they report that, in the total extraction, DepOE presented accuracy of 68%, while ReVerb reached 52% accuracy.

Aiming to improve Open IE by expanding the syntactic scope of the phrases that express relations, Mausam et al. [26] present the system OLLIE, introducing the hybrid strategy. The system is based on bootstrapping a training set used to learn pattern templates from relations extracted by the ReVerb system. It gets the pattern templates based on the dependency path connecting the arguments and the corresponding relations. After obtaining the general patterns for relation extraction, the system applies them over the corpus, obtaining new tuples. It also uses contextual information such as attribution, signaling clausal modifiers. The authors report that the system obtains a 1.9 to 2.7 times larger area under precision yield, if this one is compared to those from ReVerb and WOE systems.

ClausIE extractor, presented in [27], uses the hybrid strategy. It separates the detection of clauses and clause types from the formation of the relation tuples. A clause is defined as a part of a sentence that expresses some coherent piece of information. Combining dependency parsing and knowledge about properties of verbs, it uses a classification method to identify arguments of a relation. ClausIE also handles a limited number of non-verb-mediated relations, such as appositions, and treats possessives, introducing an artificial verb such as 'is' or 'has' in the sentence.

The authors report tests using three different datasets, achieving higher precision and recall for ClausIE than the other extractors. Specifically, ClausIE produced 2.5 to 3.5 times more correct extractions than OLLIE, the best-performing alternative method.

Bast and Haussmann [28] return to the heuristic strategy, with a method called CSD-IE (Contextual Sentence Decomposition Information Extraction) decomposing a sentence into parts that semantically 'belong together'. Facts are obtained by identifying the (implicit or explicit) verb in each such part. Their goal is to extract what they call minimal facts. They argue that an extraction as *(Ruth Gabriel, is, daughter of the actress and writer Ana Maria Bueno)*, although accurate, is not minimal since it contains two other facts hidden in the arguments. Their approach is based on the application of rules and transformations in the deep parse tree of the sentences for context decomposition and tuple extraction. The authors compare their method with ReVerb, ClausIE, and OLLIE over two of the datasets used in [27], obtaining an average of 70% precision, compatible to ClausIE and superior to the other two, while extracting triples with smaller length.

Posteriorly, in [29], the same authors propose the use of inference rules to improve the informativeness of extracted triples in Open IE. They claim that, for information retrieval applications, some of the abstract tuples extracted by the Open IE systems are useless. The authors argue that their approach allows the increase of correct and informative tuples by 15% discarding the uninformative ones. Notice that the notion of informativeness used by Bass and Haussmann is stricter than the one proposed by Fader et al. [5]. Particularly, they consider a triple as uninformative if 'there is a more precise triple that should be extracted instead'. We argue, however, that this notion of informativeness is prone to application biases, thus it may be improper to compare different systems and applications.

There is no standardization in Open IE evaluation since current studies use different measures to evaluate their results. Banko et al. [8] validate their method calculating the number of correct triples identified. Wu and Weld [10] calculate precision, recall, and F-measure based on a gold standard. The results of Etzione et al. [7] are evaluated considering AUC measure, and Gamallo et al. [25] use accuracy to evaluate their work. For a more detailed discussion on this subject, see Section 'Evaluation in open IE'.

Among the works studied, several focus on the extraction of semantic relations between words or lexical items.

The studies on the extraction of hyponymic relations from text, pioneered by Hearst [30] and continued in [31-33], were built on a prerefined number of patterns that give origin to rules. Our proposal is also rule based, but it captures a wider spectrum of semantic relations. As in [5,7,25-27], most of these are non-named semantic relations, generally represented by verbs and lack a fine-grained categorization.

Evaluation in open IE

Open IE evaluation is strongly based on information retrieval evaluation strategies. The Open IE literature reveals the use of the following measures to evaluate the effectiveness of the extractors: precision [5,7,25-29], recall [5,7,25,29], yield [26], and AUC (area under the curve) [5,7,26,27].

The use of the yield metric in Open IE is introduced by Mausam et al. [26]. Yield (Y) is calculated by multiplying the total number of extractions by the precision. The authors argue that calculating recall is difficult for Open IE results, given the volume of relationships that are extracted. They claim that the yield value is proportional to recall, so being a practical alternative in this case.

According to Boyd et al. [34] the area under the precision-recall (PR) curve is a single number that summarizes the information in the PR curve. A PR curve is built by first plotting precision-recall pairs that are obtained using different thresholds and then plotting the points in the PR space. From that, a curve is drawn and the AUC is computed. Open IE uses the confidence score of each extraction as a threshold. For instance, ReVerb [7] and OLLIE [26] assign to each extraction a confidence score using a logistic classifier trained on random Web sentences with shallow syntactic features and ClausIE [27] takes the confidence score of the dependency parses generated by the Stanford parser as the confidence score of each extraction.

Open issues in open IE

Unlike relation extraction methods focused on a prerefined set of target relations, Open IE aims at identifying all types of relationships present in the text. The goal of an Open IE system is to extract a large number of tuples describing as many relations within the text as possible with a high precision. Until now, Open IE systems have focused on verbal phrases as indicators of relationships. From the Open IE works, we realize that the extraction of relations from text is a challenging issue that is still far from being completely addressed.

The two tools that make up the Open IE state-of-the-art, OLLIE [26] and ClausIE [27], intend to improve the previous systems' main issue: identifying the arguments of relations. Mausam et al. [26] and Del Corro and Gemulla [27] claim that most of the extraction errors are due to two problems: parser failures and inability to express the relationships in the texts into binary relations, i.e., if the relationships in the text do not involve exactly two arguments, the extracted relations turn out to be incorrect because the extractors are focused on the learning of binary relationships.

An important issue in the state-of-the-art systems ClausIE and OLLIE is that, aiming at extracting the largest number of relations from the same sentence, they lose precision. They generate triples that are near to reproductions of the text and contain arguments that are not necessarily one single noun phrase. For instance, let us consider the triples extracted from the sentences 'tax revenue is the income that is gained by governments through taxation. Just as there are different types of tax, the form in which tax revenue is collected also differs'.

ClausIE : *(types of tax, differs, " ") (tax revenue, is, the income just as there are different types of tax differs) (the income just as there are different types of tax differs, is gained, by governments through taxation) (the income just as there are different types of tax differs, is gained, by governments) (tax, is, the form) (tax revenue, is collected, also in the form) (tax revenue, is collected, in the form)*

OLLIE: *(tax revenue, is, the income that is gained by governments through taxation) (tax revenue, is collected also, differs) (tax revenue, is collected also differs in, the form)*

Focusing on the gaps in the extraction of arguments, Gamallo et al. [25] state that ReVerb and DepOE do not differ significantly in the type of problems in identifying the arguments. Altogether, 65% of relationships incorrectly extracted by ReVerb represent cases in which the heuristic arguments identification screwed. An example of mistake for ReVerb is the extraction of the triple *(I, gave, him)* from the sentence 'I gave him 15 photographs' [7].

Regarding the failure to extract the relations between the arguments, the two most significant problems are: incoherent extractions and uninformative extractions. Incoherent extractions are those that do not have a meaningful interpretation. According to Etzione et al. [7], for instance, from the sentence 'the guide contains dead links and omits sites' it would be inconsistent to extract the relation *contains omits*. An example of an uninformative relation would be *is* in place of *is the author of*.

One issue in the Open IE literature is the discrepancy of the evaluation measures and results. Illustrating this situation, Etzione et al. [7] report their results using the area under precision and recall curve. Gamallo et al. [25] report precision and bar graphics representing precision, recall, and F-score. Mausam et al. [26] work with the area under precision and yield curve. Finally, Del Corro and Gemulla [27] present results using a precision $\times$ number of extractions graphic and a table with the number of correct extractions/total number of extractions.

According to Etzione et al. [7], there are three key points that must be addressed to improve the results of Open IE techniques:

1. **Extracting *n*-ary relations**, since not all relationships expressed in a text are binary. For example, from the sentence 'The first commercial airline flight was from St. Petersburg to Tampa in 1914,' we can learn two or three triples from the relational sentences as: *(the first commercial airline flight, was from, St. Petersburg), (the first commercial airline flight, was to, Tampa), (the first commercial airline flight, was in, 1914)* [25]. Trying to address this issue, Akbik and LÃűser [35] present an initial stage work performing the extraction of *n*-ary relations based on heuristic rules applied to sentences POS tagged and dependency parsed. ClausIE [27] also handles this issue and generates *n*-ary facts.
2. **Learning relationships that are not expressed by verbs**, as the relation *(Bloomberg, is the Mayor of, Seattle)* that can be inferred from the sentence 'Seattle Mayor Bloomberg said that...' Extractions such as in the example, based on noun compounds (NCs), are difficult to obtain with high precision, given the complexity of determining the semantics embedded in noun-noun sequences. Other point that is important to address is to represent the relations present in adjective-noun pairs (ANs). For instance, the sentence 'the blue car leaves the garage' presents a relation between *car* and *garage*, but, additionally, *car* has the property of being *blue* that can be interpreted as a relation. It is interesting to notice that [36] presents a solution to address this issue.
3. **Extending Open IE systems to other languages than English**. In this sense, Gamallo et al. [25] report the extraction of relations in other languages in addition to English, although they do not present qualitative assessment of these extractions.
   In addition to the points raised by Etzione et al. [7], it would be important to minimize the number of relations containing pronouns as arguments, since they are unclear outside the context of the sentence. Thus, the use of coreference resolution would increase the number of informative relations.

Generative lexicon and qualia structure

Most work in Open IE relies on the analysis of examples, manually or by a self-supervised learning method, to discover a general form of expressing semantic relations in text. Another group of works makes use of a different approach. Based on the lexical semantics of Pustejovsky [11], we argue that the semantic information held by the qualia structure of nominals may be used to identify semantic relations in a text. To explore this understanding, we adapt the rules proposed by Cimiano and Wenderoth [12] for extracting qualia structures from the Web and apply them to the task of semantic relation extraction in an Open IE system.

In this section, we first describe the Pustejovsky's qualia theory and then present Cimiano and Wenderoth approach based on rules to extract qualia structures from the Web. This work inspired us to develop our Open IE method, presented in the next section.

Pustejovsky's work on lexical semantics

Pustejovsky, in [11], proposes a rich lexical semantic theory aiming to account for the creative use of language, from the nature of word meaning to lexical creativity [37]. In contrast to the previous enumerative approaches based on the lexicon, he formalizes a set of mechanisms to lexical semantics that focus solely on the verb as a predicate of passive arguments.

Accounting for the compositional nature of the meaning, this semantic theory aggregates four different semantic descriptions of words that together account for the generative processes of semantic interpretation of the natural language[b]: a lexical typing structure; an argument structure specifying the number and types of the arguments to a given predicate; a theory of event types which are described by the lexical units; a qualia structure describing the essential attributes of an object as defined by a lexical item.

The qualia structure of a word, in which we are more interested in this work, describes the word's meaning based on four aspects, inspired by the Aristotle's principle of opposition. The basic aspects of meaning are explained as roles:

- **Constitutive role**: it expresses the relation between an object and its constituents (what it is made of). This role describes constitutive aspects of the concept such as material, weight and component elements.
- **Formal role**: it distinguishes the object within a larger domain (what it is).
- **Telic role**: it expresses purpose and function of the object.
- **Agentive role**: it expresses factors involved in the origin or 'bringing about' of an object (how it came into being).

Pustejovsky [11] uses the concept of qualia structure to represent the attributes that constitute an object, parts, purpose and function, creation mode, etc. Qualia roles express noun basic semantic features. The noun is not only connected to other nouns by traditional lexical relationships, as meronymy and hyperonymy, but also connected to verbs [38].

For instance, the term 'book' is bound in a telic role with the predicate 'read' (to read) - a process event, relating a person and the information contained in the book - and in the agentive role to the predicate 'write' (to write). An incomplete semantic description of the lexical item 'book', extracted from [11] is shown below.

$$\left[\begin{array}{l}\textbf{book}\\ \text{ARGST} = \left[\begin{array}{l}\text{ARG1} = y : \text{information}\\ \text{ARG2} = x : \text{physical\_obj}\end{array}\right]\\ \text{Qualia} = \left[\begin{array}{l}\text{FORM} = \text{hold}(x,y)\\ \text{TELIC} = \text{read}(e,w,x,y)\\ \text{AGENT} = \text{write}(e',v,x,y)\end{array}\right]\end{array}\right]$$

Pustejovsky et al. [37] argue that this semantic framework provides a different perspective when regarding many NLP questions and, such a semantic description of the lexicon should be within the very core of NLP systems. In fact, it has been served as support to many different tasks in NLP such as analysis of compounds [39] and reference resolution [40].

Cimiano and Wenderoth's qualia-based lexicons

One limiting aspect in the application of qualia-based lexicons is the necessity of building such resources, usually manually, in a time-consuming and costly process. Cimiano and Wenderoth [12] attack this problem learning the qualia structure of nouns automatically from the Web, based on lexical-syntactic patterns.

In Cimiano and Wenderoth's perception [12] of the qualia structure, the constitutive role characterizes parts or components of the object described by the noun, the formal role describes the hyperonymic relations of the noun, while the agentive and telic roles are represented by verbs that describe, respectively, an action that brings about the object in view and the purpose of this object.

Notice that, in Cimiano and Wenderoth's work, both argument structure and the predication of the verbs in the agentive and telic roles, which have an important place in Pustejovsky's [11] semantic theory, are neglected. It is important to evoque, however, that this work aims to build an auxiliary resource for lexicographers to perform the construction of lexicons, making those reasonable limitations.

We believe that the qualia structure of a nominal is an important source of information to be explored within the area of information extraction by providing both the semantic relations between concepts represented in the lexicon, as well as clues to detect the instances of such relations.

This hypothesis is supported by works such as in [39], which uses the qualia structure of nouns to interpret semantic relations within noun compounds. Although this mentioned work deals with the different problem of noun compound interpretation, it points up that the semantic information provided by the qualia structure of nominals may be used to understand the semantic relations within them.

Methods

In this section, we present an approach to extraction of relations based on pattern matching, intending to demonstrate the feasibility of an open extractor based on simple lexical-syntactic patterns. The patterns used to perform the extraction are specified in form of regular expressions as described in sections 'Qualia-based patterns' and 'Generic patterns'. We use the Penn Treebank tag set [41] in their description.

Most of the current Open IE methods depend upon pre-processing and require several steps of labeling, sometimes sophisticated. The input for our method, POS-tagged texts, is simpler to provide. Although our method requires low amount of linguistic annotation, it is heavily principled by a theory of lexical semantics to explain its adequacy, differently from most techniques in the literature.

The use of a pattern-matching (rule based) technique instead of machine learning stands on Open IE nature that is essentially linked to the Web and deals with large amounts of information, so that computational performance (in terms of speed) is mandatory. On the other hand, machine learning applications, especially the supervised ones, require the use of training data which are usually unavailable or time consuming and costly to obtain.

Like Cimiano and Wenderoth [12], we believe that semantic relations can be learned by matching lexical-syntactic. The proposed method takes a POS-tagged text as input and returns a set of triples (*arg1, rel, arg2*) describing binary relations within the text. For example, from the POS-tagged sentence 'Aristotle NN was VBD born VBN in IN Stageira NN,' the extractor generates the triple *(Aristotle,was born,Stageira*), where *arg1* = *Aristotle*, *rel* = *was born*, and *arg2* = *Stageira*.

To capture the relations in texts, we introduce 16 generic patterns to identify non-specific relations, based on the sentence structure. We also use patterns from Cimiano and Wenderoth proposal [12] to learn qualia structure, totaling three patterns to recognize qualia formal role and eight patterns based on the constitutive role (Section 'Implementation').

Qualia-based patterns

Table 1 presents the three patterns used to extract relationships based on qualia formal role, as claimed by Cimiano and Wenderoth [12]. These patterns extract hyponymy relations like *(Canada, is-a, country)*. For each pattern in the table, we present an example consisting of a sentence and a respective triple extracted using the pattern.

Table 1 Patterns used to recognize qualia formal role proposed in [12]

Example	Pattern	Triple
...works by such authors as Herrick, Goldsmith, and Shakespeare.	SUCH NP AS (NP,)*(OR|AND) NP	(Shakespeare, is a, author)
Bruises, wounds, or other injuries...	NP (,NP)* (,)? (OR|AND) (OTHER | ANOTHER) NP	(wound, is a, injury)
All common-law countries, including Canada and England...	NP (,)? (INCLUDING | ESPECIALLY) (NP,)* (OR | AND) NP	(Canada, is a, country)

Table 2 shows the eight patterns built on qualia's constitutive role. These patterns extract *made of* relations like *(ring, made of, gold)*[c]. As in Table 1, for each pattern, we present a sentence and a triple demonstrating the patterns operation.

Generic patterns

We aim at learning not only predefined but also non-specified relationships. For this task, we propose the use of new generic lexical-syntactic patterns based on the sentence structure. These patterns and corresponding examples are shown in Table 3. As in the previous tables, for each pattern, we present a sentence and a triple demonstrating the pattern operation. For instance, pattern (NP ('THAT'|'WHICH') (DT)) VB ((IN)? (WORD DT)? NP) extracts the triple *(republicans, that are from,Alabama)* from the sentence 'republicans that are from Alabama'.

Prototyping

In order to perform the evaluation, a prototype of the method was developed in JAVA, named LSOE (Lexical-Syntactic pattern-based Open Extractor) being available at https://sites.google.com/site/clarissacastella/nlp-tools.

A set of experiments was put forward. The type of parsing and the toolkit used to build the input to each system extraction rules is presented in Table 4.

Implementation

Before running LSOE, the sentences must be converted into plain text then POS tagged using OpenNLP [42] with the Penn Treebank tag set [41]. These POS-tagged sentences will be used as input for LSOE.

The patterns presented in Tables 1, 2 and 3 are implemented using regular expressions. We apply a simple noun phrase (NP) identification method based on the application of the following regular expressions:

$$\begin{aligned} \text{SUB1} &= [\,/-\text{A}-\text{Z0}-9]*(\text{NNS} \mid \text{NN} \mid \text{NNP}) \\ \text{SUB2} &= [\,/-\text{A}-\text{Z0}-9]*(\text{NN} \mid \text{NNS} \mid \text{JJ} \mid \text{NNP}) \\ &\quad\ [\,/-\text{A}-\text{Z0}-9]*(\text{NN} \mid \text{NNS} \mid \text{NNP}) \\ \text{SUB3} &= \text{SUB1 (OF IN) SUB1} \\ \text{NP} &= (\text{SUB1}|\text{SUB2}|\text{SUB3}) \end{aligned}$$

Results and discussion

In this section, we compare the prototype performance with ReVerb and DepOE. We also discuss LSOE results providing an error analysis.

We have performed two evaluation rounds, each one using a different input corpus. The input of the first round was a corpus of 217 randomly selected sentences from Wikipedia articles related to the Philosophy of Language domain. The second round used a domain-independent corpus containing 2,701 Wikipedia articles from Wikicorpus [43].

We expect that LSOE obtains precision compatible with rule-based Open IE systems and that it extracts relations that are not learned by them. We also calculate yield as proposed in [26] for each system measuring the number of new relations obtained.

Evaluation setup

The evaluation of the system was carried by a manual assessment of the results. We compared LSOE with the

Table 2 Lexical-syntactic patterns based on qualia constitutive role used in [12]

Example	Pattern	Triple
A ring is made up of gold	NP IS (MADE UP OF) NP	(ring, made of, gold)
Rings are made up of gold	NP (ARE MADE UP OF) NP	(rings, made of, gold)
Ring is made of gold	NP (IS MADE OF) NP	(ring, made of, gold)
Rings are made of gold	NP (ARE MADE OF) NP	(rings, made of, gold)
The whole comprises the parts	NP (COMPRISES) NP	(whole, comprise, parts)
The members comprise the team	NP (COMPRISE) NP	(members, comprise, team)
The package consists of brochures	NP (CONSISTS OF) NP	(package, consists of, brochures)
The lands consist of valleys	NP (CONSIST OF) NP	(lands, consist of, valleys)

Table 3 Lexical-syntactic patterns used to identify non-specified relations in texts

Example	Pattern	Triple
A central branch of metaphysics is ontology;	'A' NP 'OF' NP 'IS' NP	(ontology, is a central branch, metaphysics)
Biology is the study of living things	NP 'IS' 'THE' EXP 'OF' NP	(Biology, study, living things)
Aristotle born in Stageira	NP VB ('IN' | 'AT') NP	(Aristotle, born, Stageira)
Aristotle was born in Stageira	NP ('WAS' | 'IS') VP ('IN' | 'AT') NP	(Aristotle, born, Stageira)

two other rule-based open extractors: ReVerb and DepOE systems. The type of parsing and the tools used in this task by each extractor is presented in Table 5.

After running each system over the same input sentences, two human judges evaluated the triples generated by the systems as correct or incorrect, following the same procedure from [7]. Uninformative or incoherent triples were classified as incorrect. According to Etzioni et al. [7], incoherent extractions are 'the cases where extracted relation phrase has no meaningful interpretation,' while uninformative extractions are those which 'omit critical information.' Only results from the subset of the data where the judges agree were considered for results evaluation.

For instance, given the sentence 'Kantianism is the philosophy of Immanuel Kant', LSOE generated the triple *(Kantianism,is the philosophy of, Immanuel Kant)* and DepOE the triple *(Kantianism, is, the philosophy of Immanuel Kant)*, both labeled as correct. In contrast, ReVerb generated the triple *(Kantianism, is, the philosophy)*, labeled as incorrect due to being uninformative.

First round

In the first evaluation round, over the Philosophy of Language domain, the judges reached agreement on 62% of the judgments. Table 4 gives an overview of the results of this evaluation round. It shows the number of triples extracted by each system and their classification as correct or incorrect. From the 311 tuples extracted by the LSOE in the Philosophy of Language, 169 (approximately 54%) were extracted by the rules based on the qualia structure. Figure 1 shows that LSOE achieves both higher precision (43%) than ReVerb (8%) and DepOE (26%) and higher yield, 133.73 for LSOE against 8.2 for ReVerb and 31.72 for DepOE, in this domain-specific context.

Regarding the relation phrases inside the tuples, LSOE learned 119 different types, as presented in Table 6. As expected, most relation description came from the generic patterns, since the qualia-based patterns extract a pre-established number of relations, namely, only two (*is_a* and *consists*) relation phrases were extracted using Cimiano and Wenderoth [12] patterns.

Second round

In the second evaluation round, using the Wiki corpus input, the judges reached an agreement in 68% of the triples. This time, since the input is much larger than the first one, the number of triples obtained by each system significantly increased. LSOE generated 2,539 triples, while DepOE extracted 4,279 triples and ReVerb extracted 52,739 triples, from which 26,740 have confidence rate equal or greater than 70%.

Figure 2 shows that the systems achieve their better results using a larger input corpus. The precision obtained in this evaluation was 54% for LSOE, 49% for ReVerb, and 27% for DepOE. It is interesting to note that ReVerb performed exceptionally better in this context, obtaining better outcome than DepOE. LSOE continues to obtain greater precision than the other two systems, however the difference between LSOE's and ReVerb's performance was much smaller in this round of evaluation.

Regarding the yield metric, LSOE obtained 1,371.06, while DepOE obtained 115.33 and ReVerb 13842.11. Notice that such a high yield measure for the ReVerb system comes from the sheer amount of relations it extracted from the text (considering only those with confidence over 70%). Even though DepOE extracted around 70% more triples than LSOE, given LSOE's high precision, it outperformed DepOE in regard to the yield metric.

Regarding the lexical-syntactic patterns that LSOE used to identify relations, some interesting considerations could be brought. We observed that 29% of the tuples were obtained from the rules based on qualia structure. From this subset, 71% were assessed as correct and 29%

Table 4 Number of correct and incorrect triples extracted in the first evaluation round

	LSOE	ReVerb	DepOE
Correct	133 (43%)	8 (10%)	27 (26%)
Incorrect	178 (57%)	74 (90%)	95 (92%)
Total	311	82	103

Table 5 The evaluated systems input

System	Input	Toolkit
ReVerb	POS-tagged and NP-chunked text	OpenNLP
DepOE	POS-tagged and dependency-parsed texts	TreeTagger and DepPattern
LSOE	POS-tagged text	OpenNLP

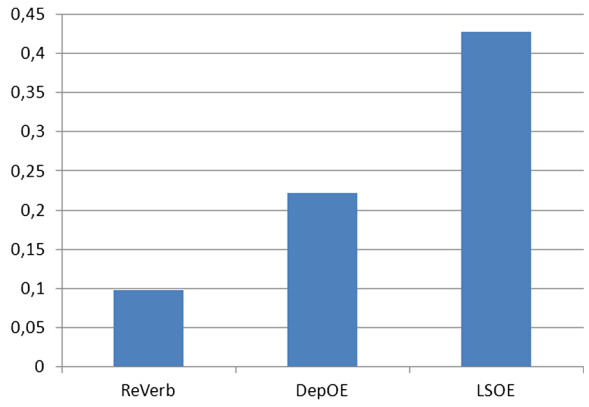

Figure 1 Precision for ReVerb, DepOE, and LSOE in the first evaluation round. A bar chart of the precision achieved in the first round of evaluation by the systems ReVerb, DepOE, and LSOE showing the ReVerb with the lower precision (around 0.1) followed by DepOE (around 0.2) and LSOE with the best result (around 0.4).

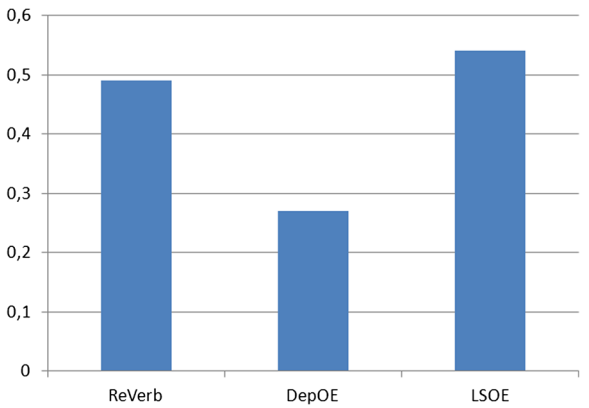

Figure 2 Precision for ReVerb, DepOE, and LSOE in the second evaluation round. A bar chart of the precision achieved in the second round of evaluation by the systems ReVerb, DepOE, and LSOE. In this round, LSOE obtains better precision than DepOE and just a little greater precision than ReVerb. It shows the DepOE system with lower precision (around 0.3) followed by ReVerb system (around 0.5) and LSOE with a little higher result than ReVerb (around 0.55).

as incorrect. Most of the extracted relations, i.e., the other 71%, were obtained using the generic rules proposed in Table 3. From this subset, 75% were evaluated as correct and 25% as incorrect.

Table 7 shows the number of relations inside and outside the intersection of the sets of relations extracted by LSOE, ReVerb, and DepOE. In the first round, LSOE learned 117 relations that were not learned by ReVerb and DepOE. In the second round, LSOE learned 541 relations that were not learned by ReVerb and 531 relations not learned by DepOE. Observing this data, we realize that LSOE performance can be taken as complementary to the other extractors, since the relations extracted by each method little recur.

Regarding the relation phrases inside the tuples, LSOE learned 871 different types in the second round, as presented in Table 6. As in the first round, only two (*is_a* and *consists*) relation phrases were extracted using Cimiano and Wenderoth [12] patterns. In the evaluation set, 35% of the tuples were obtained using qualia-based rules and 65% generic rules. From those, 26% of the tuples obtained by generic patterns were judged as incorrect and 35% as correct. Regarding the relations obtained with Qualia-based patterns, 59% were judged as correct and 41% as incorrect.

Overall perception on the evaluation rounds

Concerning the number of relations that appear in the triples generated by LSOE, as shown in Table 6, there is a similar behavior between the two rounds. That is, regardless the domain and the size of the input corpus, the prototype identified a similar proportion between different relations and extracted triples. These are initial evaluation rounds of the proposed method, and further tests are needed to better explain the disagreement between the two rounds and the three systems. A general analysis of the results indicates the potential of the proposed method.

The relationship identified more often by LSOE was the subsumption or instantiation relation (*is-a*) with 37 correctly identified instances. Similar relations identified by the other two systems (*is, are, was, were*) account for nine instances for the ReVerb system and 24 instances for the DepOE system. Note, however, that, especially in the case of DepOE, some of these relations may not be regarded as a subsumption or instantiation. For instance, from the sentence 'Some notable leaders were Ahmed Ullah [...]', DepOE system identified the relation *(Some notable leaders, were, Ahmed Ullah)*. The roles of the arguments are reversed in the triple, so that this relation could never be understood as an instantiation.

Table 6 Comparison between assessment rounds: triples and relation phrases

	First round	Second round
Triples	311	2,359
Relation phrases	119	871

Table 7 Intersection of relations extracted: relations generated by LSOE that were generated or not by ReVerb and DepOE in each evaluation round

	ReVerb		DepOE	
	Repeated	Not repeated	Repeated	Not repeated
First round	2	117	2	117
Second round	330	541	340	531

Discussion

Regarding the performance of the systems in the first round, LSOE had much better results, both in precision and yield, than the other two systems. We identify two main reasons for this fact:

1. From our perception of the extractors as a whole, ReVerb and DepOE were designed to work with a much larger and multiple domain data entry. From that, we conclude that the small number of sentences used as input did not allow them to accomplish their best performance.
2. The nature of the texts in the domain-specific corpus was formal and academic. Qualia-based patterns are very powerful to identify a great number of triples in this kind of sentences, and LSOE may benefit from this.

In the second round of evaluation, the other two systems achieved much better performance (c.f. Figure 2). From the 150 triples in the second round analysis, LSOE identified 95 different relations, while ReVerb and DepOE identified 138 and 113 relations, respectively.

Regarding the yield metric, ReVerb obtained a much higher value than the other two extractors followed by distance from LSOE and DepOE that besides extracting around 70% more triples than LSOE, given LSOE's high precision, outperformed DepOE. Since ReVerb extracted much more tuples than LSOE and achieved 49% of precision, while LSOE achieved 54%, it is clear why ReVerb outcomes LSOE in this metric.

Concerning the improvement in the performance of LSOE from the first to the second round of evaluation, we identify two main reasons:

- A common mistake of the first round was the interpretation of relations expressed that match the patterns used to recognize qualia formal role. For instance, from the sentence 'Members of the principally british associationist school, including John Locke, David Hume, James Mill, John Stuart Mill and Alexander Bain [...],' LSOE originates the triples *(John_Locke, is-a, school), (David_Hume,is-a,school), (James_Mill, is-a, school), (John_Stuart_Mill, is-a, school)*, and *(Alexander_Ba, is-a, school)* when a better extraction would be *([Philosopher name], member_of, the principally british associationist school)*.
- In the academic domain of Philosophy of Language, the generic patterns (which were responsible for 46% of extractions) are less effective and the extractions are prone to errors.

Unfortunately, the set of relations extracted in both rounds cannot be directly or automatically compared, since the different systems extract the same relations expressed in the sentences in different formats, as in the previous example about Kantianism (section 'Evaluation setup'). It is important to emphasize that for both evaluation input sets, LSOE and ReVerb generated tuples more related to the standard *(concept - relates to - concept)*, as usually seen in ontologies and other knowledge structures, while DepOE's tuples reflected the textual relations in the sentences, usually comprising an entire sentence or large chunks of them, as presented in the second line of Table 8.

To better quantify the differences of the extractors behavior, we measured the mean length of the relation arguments for the three different systems. We observe, as noted earlier, that both ReVerb and LSOE have similar behavior with a mean length of approximately 11 and 12 characters by argument, respectively, while DepOE extracted relations with longer arguments, with mean length of approximately 35 characters, such as the triple *(descartes the foundations of solipsism, are in, turn the foundations of the view that the individual's understanding of any and all psychological concepts)*. On the other hand, from the sentence 'Aristotle was born in Stageira, Chalcidice, in 384 bc, about east of modern-day Thessaloniki,' LSOE and ReVerb extracted the triples *(Aristotle, was_born, Stageira)* and *(Aristotle, was born in, Stageira)*, respectively, while DepOE did not extract any triple. In contrast, from the sentence 'Descartes: the foundations of solipsism are in turn the foundations of the view that the individual's understanding of any and all psychological concepts (thinking, willing, perceiving, etc.)'.

The systems seem to be quite different in their approach to extract relations in text, since it was not easy to find examples of relations extracted by the three systems from the same sentence. One of the few examples are the tuples *(formal logic, is, the study), (formal_logic, is_the_study_of, inference)*, and *(formal logic, is, the study of inference with purely formal content)*, obtained by ReVerb, LSOE, and DepOE, respectively, from the sentence 'formal logic is the study of inference with in purely formal content.'

Regarding the incorrect tuples, they can be classified as incoherent or uninformative. From the tuples evaluated as incorrect in the second round, 8% were considered incoherent and 93% uninformative for DepOE, 11% incoherent and 89% uninformative for ReVerb, and 48% incoherent and 52% uninformative for LSOE. We can see that DepOE and ReVerb generated few incoherent and more

Table 8 Examples of triples extracted by the three systems

System	Triple
DepOE	*(an intension, is, any property or quality connoted by a word)*
ReVerb	*(Allan Gotthelf, is a, writers)*
LSOE	*(Kuba_Saeed, is-a, district)*

uninformative tuples. Thus, most of the errors of these two extractors were due relations where at least one of the arguments was incomplete generated. For example, DepOE generated uninformative relation *(Billings Bridge, is, 5 minutes)* from the sentence 'To the south, Billings Bridge is 5 minutes away, and the Airport&Station is 18 minutes away' and the ReVerb the uninformative relation *(Sonia Gandhi, has been described as, India 's best)* from the sentence 'Sonia Gandhi has been described as India 's best dressed politican.' On the other hand, LSOE produced a similar number of uninformative and incoherent relations. This means that LSOE patterns may extract relations that make no sense but obtain less incomplete ones. For example, LSOE extracted the incoherent relation *(play, was written, part)* from the sentence 'the play was written in part in response to the events in September 11.' As in the example, the error behind the inconsistent relations is in the relation part of the tuple, when the pattern cuts off one or more words from one of the items of the tuple (in this case, the second argument would be *in part in response* instead of *part*).

We notice that 22% of the errors made by LSOE in the second round occurred in choosing the noun phrases that compose the arguments of the relations. For example, in the sentence segment below, our method extracted the triple *(Japan, is-a, East Asia)* instead of *(Japan, is-a, countries of East Asia)*, by choosing the smaller noun phrase fitting the pattern.

> *[NP, The countries of [NP,East Asia]] including China, Japan And Korea as well as Vietnam [...]*

Regarding the use of the patterns, the one that generates the largest number of erroneous triples was: NP (,)? (INCLUDING | ESPECIALLY) (NP,)* (OR | AND) NP.

Other qualia-based pattern occurrences were relatively rare. This is not surprising considering the size of the corpus. According to [12], it is well known that patterns as Hearst's for hiponymy detection occur rarely in text. Cimiano and Wenderoth [12] suggest that applying rules that search those patterns in a large corpora, as the Web, would greatly increase the results.

Notice that the few errors observed which are related to qualia-based patterns seem to fall in the same problem as above, e.g., our approach identified the triple *(world, consists of, nothing)* processing the following sentence, while more suitable choices would be *(world, consists of, nothing but objective particles in fields of force)* or *(world, consists of, objective particles in fields of force).*

> *[...] the world consists of nothing but objective particles in fields of force [...]*

Another issue in LSOE performance was the treatment of quotation and indirect discourse. For example, the pattern NP IS THE EXP OF NP extracts the relation *(X, is the king of, France)* in the following sentence segment:

> *[...]the above can be expressed in a more strict logical form (where K(X) means 'X is the king of France[...]*

Clearly, *(X, is the king of, France)* should not be identified as information expressed in the text, with *'X is the king of France'* being inside quotation marks. Another example is the triple *(world, is, God)* identified in the following sentence, when a much better suited candidate would be *(Pantheism, can be summed up as, The world is in God and God is in the world, but God is more than the world and is not synonymous with the world).*

> *Pantheism can be summed up as 'The world is in God and God is in the world, but God is more than the world and is not synonymous with the world.'*

The set of qualia-based patterns is yet very limited. We implemented only patterns for the formal and constitutive roles, comprising 11 patterns in total, some of them of very limited application - especially the ones for the constitutive role. Other patterns may be included in this set, such as those from Girju et al.'s [44] work on meronymy relations, increasing the coverage of such patterns.

We believe that LSOE performance can be enhanced by refining the lexical-syntactic patterns used in the relation identification, improving the recognition of the arguments of a relation, and using parse trees. Furthermore, named entity recognizers can be applied to improve recognition of nouns.

Also, it would be useful to treat relations containing pronouns as arguments through the use of coreference resolution. Since these relations are unclear in some contexts, coreference resolution increases the number of informative relations.

Overall, the results produced by our method are encouraging when compared to rule-based Open IE methods. In a domain-independent corpus, our method extracts fewer triples but achieve compatible precision to DepOE and ReVerb. To overcome this coverage issue, we suggest the inclusion of new generic rules. Notice, however, that newer methods in OpenIE, such as OLLIE [26] and ClausIE [27] that belong to the so-called second-generation OpenIE, are much more dependent both on syntactic information and machine learning to avoid errors or uninformative extractions. We believe that our work can be extended to apply classifiers as a post-processing method to filter the current system extractions.

Conclusions

In this work, we have presented the Open IE paradigm of relation extraction from texts, its importance, and challenges. In order to go further this study, we have developed an Open IE method based on lexical-syntactic patterns. The key idea is to provide a simple solution to perform rule-based extraction of triples using POS-tagged text. The extractor identifies relationships by applying lexical-syntactic patterns based on Pustejovsky's qualia structure [11], as proposed by Cimiano and Wenderoth [12], and generic patterns that identify non-specified relationships based on the sentence structure.

We conducted an initial evaluation of the method by building a prototype and comparing its performance with ReVerb and DepOE. The results demonstrate that LSOE extracts relations that are not learned by the other extractors, achieves compatible precision, and needs to improve its yield for input from multiple domains.

In the future, we aim at improving the performance of LSOE with the extraction of relations expressed by verbs by including new lexical-syntactic patterns, implementing post-processing rules and other methods based on syntactic information, as dependency parsing. The discussion in Section 'Discussion' also suggests that the improvement in the argument identification is crucial to produce better results. Most of the errors in the extracted triples came from situations where the relation was correctly identified but not one of the arguments.

We are also working in a novel approach to address the challenge of automatic extraction of relations described within NCs and ANs using Open IE paradigm [36]. Etzione et al. [7] indicate that learning relationships that are not expressed by verbs is a key gap in current Open IE state-of-the-art methods. Since verbs are not the only way to express relations between nouns in a text, it is important to create alternatives to learn those relations. For example, the noun compounds *cheese knife* and *kitchen knife* inform the relations (*knife, for cutting, cheese*) and (*knife, used in, kitchen*), respectively. Until this point, we are not aware of any other work beyond ours that perform this task.

Endnotes

[a] A CRF (conditional random fields) classifier identifies to which of a set of categories a new information belongs, based on CRF, a probabilistic framework for labeling and segmenting structured data [45].

[b] Note that this framework of semantic descriptions as detailed in [37] is turned to nouns, and it does not account for discourse and pragmatic factors.

[c] It is interesting to point out that the authors propose different patterns to deal with singular and plural forms of verbs, instead of dealing with lemma information having more generic patterns.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CCX envisioned and formalized the method, contributed to the theoretical review, implemented the prototype, performed experiments and analyzed results. MS contributed to the theoretical review, performed experiments and analyzed results. VLSL supervised all methodological and experimental aspects of the research. All authors read and approved the final manuscript.

Author details
[1] Faculty of Informatics, Pontifical Catholic University of Rio Grande do Sul, Av. Ipiranga 6681, Porto Alegre, Brazil. [2] Institute of Informatics, Federal University of Rio Grande do Sul, Av. Bento Gonçalves, 9500, Porto Alegre, Brazil.

Received: 30 June 2014 Accepted: 8 April 2015
Published online: 06 May 2015

References

1. Nakashole NT (2012) Automatic extraction of facts, relations, and entities for web-scale knowledge base population. PhD thesis, University of Saarland
2. Ruiz-Casado M, Alfonseca E, Okumura M, Castells P (2008) Information extraction and semantic annotation of wikipedia. In: Proceedings of the 2008 conference on ontology learning and population: bridging the gap between text and knowledge. IOS Press, Amsterdam, The Netherlands. pp 145–169
3. Angeli G, Manning C (2013) Philosophers are mortal: inferring the truth of unseen facts. In: Proceedings of the seventeenth conference on computational natural language learning. Association for Computational Linguistics, Sofia, Bulgaria. pp 133–142
4. Huang J, Etzioni O, Zettlemoyer L, Clark K, Lee C (2012) RevMiner: an extractive interface for navigating reviews on a smartphone. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology. UIST '12, ACM, New York. pp 3–12
5. Fader A, Soderland S, Etzioni O (2011) Identifying relations for open information extraction. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. pp 1535–1545
6. Eichler K, Hemsen H, Günter N (2008) Unsupervised relation extraction from web documents. In: Proceedings of the international conference on language resources and evaluation. European Language Resources Association (ELRA), Marrakech, Morocco
7. Etzioni O, Fader A, Christensen J, Soderland S, Mausam M (2011) Open information extraction: the second generation. In: Proceedings of the twenty-second international joint conference on artificial intelligence, vol. 1. AAAI Press, Barcelona, Spain. pp 3–10
8. Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O (2007) Open information extraction from the web. In: Proceedings of the 20th international joint conference on artificial intelligence. Morgan Kaufmann Publishers Inc., Hyderabad, India
9. Li H, Bollegala D, Matsuo Y, Ishizuka M (2011) Using graph based method to improve bootstrapping relation extraction. In: Computational linguistics and intelligent text processing, vol. 2. Springer, Berlin. pp 127–138
10. Wu F, Weld DS (2010) Open information extraction using Wikipedia. In: Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA. pp 118–127
11. Pustejovsky J (1995) The generative lexicon. The MIT Press, Cambridge
12. Cimiano P, Wenderoth J (2005) Automatically learning qualia structures from the web. In: Proceedings of the ACL-SIGLEX workshop on deep lexical acquisition. Association for Computational Linguistics, Ann Arbor, Michigan. pp 28–37
13. Nastase V, Nakov P, Séaghdha DÓ, Szpakowicz S (2013) Semantic relations between nominals. Synth Lect Hum Lang Technol 6(1):1–119
14. Khoo C, Na JC (2006) Semantic relations in information science. Annu Rev Inf Sci Technol 40:157–228
15. Murphy ML (2003) Semantic relations and the lexicon. Cambridge University Press, Cambridge

16. Banko M, Etzioni O (2008) The tradeoffs between open and traditional relation extraction. In: Proceedings of the 46th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Columbus, Ohio, USA. pp 28–36
17. DeJong GF (1979) Prediction and substantiation: a new approach to natural language processing. Cognitive Sci 3:251–273
18. DeJong GF (1982) An overview of the frump system. In: Strategies for natural language processing. Lawrence Erlbaum, Hillsdale. pp 149–176
19. Cowie J, Lehnert W (1996) Information extraction. Commun ACM 39(1):80–91
20. Banko M (2009) Open information extraction for the web. PhD thesis, University of Washington
21. Min B, Shi S, Grishman R, Lin C-Y (2012) Towards large-scale unsupervised relation extraction from the web. Int J Semantic Web Inf Syst (IJSWIS) 8(3):1–23
22. Yates A, Etzioni O (2007) Unsupervised resolution of objects and relations on the web. In: HLT-NAACL. Association for Computational Linguistics, Vancouver, British Columbia, Canada. pp 121–130
23. Kok S, Domingos P (2008) Extracting semantic networks from text via relational clustering. In: Machine Learning and Knowledge Discovery in Databases. Springer, Berlin. pp 624–639
24. Soderland S, Mandhani B (2007) Moving from textual relations to ontologized relations. In: AAAI spring symposium: machine reading. AAAI Press, Palo Alto, California, USA. pp 85–90
25. Gamallo P, Garcia M, Fernández-Lanza S (2012) Dependency-based open information extraction. In: Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP. Association for Computational Linguistics, Avignon, France. pp 10–18
26. Mausam, Schmitz M, Bart R, Soderland S, Etzioni O (2012) Open language learning for information extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, Jeju Island, Korea. pp 523–534
27. Del Corro L, Gemulla R (2013) Clausie: clause-based open information extraction. In: Proceedings of the 22nd international conference on world wide web. International World Wide Web Conferences Steering Committee, Rio de Janeiro, Brazil. pp 355–366
28. Bast H, Haussmann E (2013) Open information extraction via contextual sentence decomposition. In: ICSC. IEEE, Irvine, California, USA. pp 154–159
29. Bast H, Haussmann E (2014) More informative open information extraction via simple inference. In: Advances in information retrieval. Springer, Berlin. pp 585–590
30. Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on computational linguistics, vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA. pp 539–545
31. Berland M, Charniak E (1999) Finding parts in very large corpora. In: Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics. Association for Computational Linguistics, College Park, Maryland, USA. pp 57–64
32. Rosario B, Hearst M (2001) Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: Proceedings of the 2001 conference on empirical methods in natural language processing. Association for Computational Linguistics, Pittsburgh, PA, USA. pp 82–90
33. Girju R, Moldovan DI (2002) Text mining for causal relations. In: Proceedings of the fifteenth international florida artificial intelligence research society conference. AAAI Press, Cambridge. pp 360–364
34. Boyd K, Eng K, Page CD (2013) Area under the precision-recall curve: point estimates and confidence intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F (eds). Machine learning and knowledge discovery in databases. Lecture notes in computer science vol. 8190. Springer, Berlin. pp 451–466
35. Akbik A, Löser A (2012) Kraken: N-ary facts in open information extraction. In: Proceedings of the NAACL-HLT 2012 joint workshop on automatic knowledge base construction and web-scale knowledge extraction. Association for Computational Linguistics, Montreal, Canada. pp 52–56
36. Xavier C, Lima VS (2014) Boosting open information extraction with noun-based relations. In: Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland. pp 96–100
37. Pustejovsky J, Havasi C, Littman J, Rumshisky A, Verhagen M (2006) Towards a generative lexical resource: the brandeis semantic ontology. In: Proceedings of the fifth language resource and evaluation conference. European Language Resources Association (ELRA), Genoa, Italy Vol. 7. pp 1702–1705
38. Bouillon P, Claveau V, Fabre C, P S (2002) Acquisition of qualia elements from corpora - evaluation of a symbolic learning method. In: Proceedings of the third international conference on language resources and evaluation. European Language Resources Association (ELRA), Las Palmas, Canary Islands, Spain. pp 208–215
39. Johnston M, Busa F (1996) Qualia structure and the compositional interpretation of compounds. In: Proceedings of the ACL SIGLEX workshop on breadth and depth of semantic lexicons. Association for Computational Linguistics, Santa Cruz, California, USA. pp 77–88
40. Bos J, Buitelaar P, Mineur A-M (1995) Bridging as coercive accommodation. In: Klein E, Manandhar S, Nutt W, Siekmann J (eds). Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95). Computerlinguistik an der Universität des Saarlandes, Edinburgh, UK. pp 1–16
41. Marcus MP, Marcinkiewicz MA, Santorini B (1993) Building a large annotated corpus of English: the penn treebank. Comput linguist 19(2):313–330
42. Baldridge J (2005) The openNLP maximum entropy package. Technical report, Apache Software foundation
43. Reese S, Boleda G, Cuadros M, Padró L, Rigau G (2010) Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In: Proceedings of the seventh international conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), La Valleta, Malta. pp 1418–1421
44. Girju R, Badulescu A, Moldovan D (2003) Learning semantic constraints for the automatic discovery of part-whole relations. In: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1. Association for Computational Linguistics, Edmonton, Canada. pp 1–8
45. Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth International Conference on Machine Learning (ICML'01). Morgan Kaufmann, San Francisco. pp 282–289