J Braz Comput Soc (2013) 19:553–571
DOI 10.1007/s13173-013-0116-8

SURVEY PAPER

A review on Relation Extraction with an eye on Portuguese

Sandra Collovini de Abreu · Tiago Luis Bonamigo · Renata Vieira

Received: 4 July 2012 / Accepted: 12 June 2013 / Published online: 9 July 2013

Abstract The task of Relation Extraction is one of the main challenges in Natural Language Processing. We present a review of the state-of-the-art for Relation Extraction in free texts, addressing the progress and difficulties of the area, and situating Portuguese in that frame. We discuss the different aspects related to this task, considering the main computational strategies, used resources, as well as the evaluation methods applied. We also give special attention to the literature for Portuguese tools, which need further progress. On the best of our knowledge, this is the first comprehensive survey on Relation Extraction to include the state of the work done for Portuguese.

Keywords Information Extraction · Relation Extraction · Semantic Relations · Named Entity Recognition · Natural Language Processing · Portuguese

1 Introduction

Information Extraction (IE) from text has been extensively studied in various research communities including Natural Language Processing (NLP), Information Retrieval (IR), Web mining, and others. It has applications in a wide range of domains; the specific type and structure of the information to be extracted depends on the need of the particular application [60]. A good example is the business intelligence domain, in which financial professionals often need to seek specific information from news articles to help making their everyday decisions. Another possible application is in the intelligence area, where analysts review large amounts of text to search for information, such as on people involved in terrorism events, the weapons used and the targets of the attacks. Finding such information from text automatically requires IE technologies/applications, such as Named Entity Recognition (NER) and Relation Extraction (RE).

In this context, NER task aims at identifying, disambiguating and attributing a semantic category to those Named Entities (NE) within the text, such as names of *Organization*, *Location*, *Person*, among others. A survey on this task is found in [74]. Appropriate recognition of NE within texts is of utmost importance for IE, since the meaning of the texts is usually anchored in these entities. According to [86], informative texts such as news contents usually refer to entities and specific events rather than to generic concepts.

However, many IE tasks cannot rely on NER alone, requiring also the identification of relations established among such entities. For example, considering the business domain, only relying on identifying company names contained within a news article is not as informative as identifying the relation "*is acquired by*" between two entities representing companies or the relation " *is appointed CEO of* " between entities representing *person* and *company*, respectively [17,64,89].

Given the importance of exploring relations for a more accurate understanding of language, the need for further advances in techniques for identifying and extracting them emerged, thus, the establishment of the Relation Extraction task was necessary. RE is, thus, an important IE task in NLP, and it helps many others, such as Question Answering (QA), IR, summarization, semantic Web annotation, construction and extension of lexical resources and ontologies.

S. C. de Abreu (✉) · T. L. Bonamigo · R. Vieira
PUCRS, Porto Alegre, Brazil
e-mail: sandra.abreu@acad.pucrs.br

T. L. Bonamigo
e-mail: tiago.bonamigo@acad.pucrs.br

R. Vieira
e-mail: renata.vieira@pucrs.br

The problem of Relation Extraction from natural language texts has been studied extensively, including in news articles, scientific publications, blogs, emails, and sources like Wikipedia, Twitter and the Web [89]. There is an increasing interest in Relation Extraction, mostly motivated by the exponential growth of information made available through the World Wide Web, which makes the tasks of researching and using this massive amount of data impossible through manual means. The popularization of social networking websites such as Twitter and Facebook, that are basically propelled by users inserting new data on a frequent basis, results in a daily generation of hundreds of millions of small texts. That context makes Relation Extraction an even more complex and relevant research area [35].

One of the main challenges for extracting information from the Web and make users able to use this tool properly, is that its contents do not adhere to any specific structure, which makes the automated processing of such content difficult. Web Semantics attempts to provide the Web with such structure, giving proper meaning to content so that it can be used in a more efficient manner by computer systems. In order to structure that vast amount of unstructured information, resources and techniques within the fields of Machine Learning and NLP, in particular of IE and Text Mining, are needed. As an example, ontologies can be a valuable resource to represent and provide structure to the meaning contained within Web pages through the annotation of domain-specific terms [82]. Given the usefulness of ontologies for that purpose, anything that might assist semantic Web annotation and ontology learning is a valuable resource, and Relation Extraction is an advantageous means for supporting those tasks [93].

Several approaches have been proposed for the extraction of relations from unstructured sources, such as supervised or unsupervised machine learning; corpus-based techniques; linguistic strategies; resources as lexical databases and ontologies; rule-based heuristics and hybrid systems. For some languages, such as English, there is extensive research and literature regarding RE [1,5,10,21,25,33,34,36,48,54,57,58,65,70,79,93,94,100], while for Portuguese, there are fewer references to existing work dealing with RE [11,15,38,42,85,96,101].

Considering the context of emergent economies such as the BRICS group (Brazil, Russia, India, China and South Africa) and the growing interest of foreign investors due to the opportunities in those countries, a better understanding of the local languages—and, thus, of Information Extraction—can lead to strategic advantages. The Brazilian economy growth has led the country to overtake the United Kingdom as the sixth biggest economy in the planet, leading to an increased appeal for the Portuguese language, similar to what happened to Mandarin during the accelerated growth of Chinese economy through the last decade.

Given this scenario of increased attention to Portuguese tools and the development of research towards techniques that can support this interest, this paper presents the state-of-the-art for RE in open texts and attempts in order to compare the approaches for Portuguese tools to it. To the best of our knowledge, this is the first comprehensive survey of RE to include the state of the work done for Portuguese.

This paper is organized as follows. Section 2 characterizes what is a relation, the relevant arguments involved in the scope of this paper and the corresponding NLP task of Relation Extraction. Section 3 describes the state-of-the-art systems that perform Relation Extraction and their corresponding approaches, discussing Portuguese systems as well. Section 4 presents the linguistic resources available for RE, emphasizing those that deal with the Portuguese language, and Sect. 5 discusses Relation Extraction evaluation works and joint contests, highlighting the evaluations that focus on the Portuguese language. Finally, Sect. 6 presents some concluding remarks.

2 Relation Extraction

2.1 Relations

According to Gruber [52] relations are a set of tuples that represent a relationship among objects within an object of discourse. In particular, semantic relations are relations between concepts or meanings involving different linguistic units and components, such as the "*headquartered in*" relation relates *Organization* class with *Location* class. For example, we have the relation "*headquartered in*" between "*Microsoft Corporation*" and "*Redmond*" within the sentence "*Microsoft Corporation is headquartered in Redmond*".[1]

In the literature there is a variety of relation types, and a relation is considered relevant according to several factors, mainly the type of information that is being analyzed and the objective of the extraction task. For instance, the identification of the *protein-organism-location* relations in the text of biomedical articles provide information about where a protein is located in an organism, giving a valuable clue to the biological function of the protein and helping in the diagnostic [66]. One way to define relevant relations for a certain domain or textual genre, as well as identifying patterns that describe such relations, is through data analysis. According to Hearst [55] different relations can be expressed by utilizing a small number of lexical-syntactical patterns.

Given the lack of a standard definition of target relations to be considered in the Relation Extraction tasks, Table 1 presents some of the relations defined in previous works (the related works are further discussed in Sect. 3). Among

[1] http://en.wikipedia.org/wiki/Microsoft.

Table 1 Overview of relation definitions in the literature

Relations	Works for English	Works for for Portuguese
author-title / authorOf / author_of / *work_of*	[3,10]	[15,45]
location / locatedIn / location-of / located	[17,54,73,75]	[11,15,101]
employment-organization / role / organization-role	[54,75,94,95]	[24]
quotation-author	[62,80]	[38,90]
CeoOf / person-organization	[17]	[24]
birthPlace / *bornIn*	[4,64]	[15]
living_in / *home_of* / residence / place_of	[75]	[15,45]
family_ relation / social / parent	[75]	[15,45]
member / member-of	[75]	[24]
manufacturing / manufactured_ by / produces	[17]	[15,45]
organization-headquarters / headquarteredIn / based-in	[1,17,75]	–
employee-of	[73,94]	–
acquired /acquisition / merge-acquisition	[3,4,17,54]	–
wonAward / hasWonPrize	[4,64]	–
part-of / part-whole / part / subsidiary	[54,75,79]	–
management / affiliate-partner / founder	[75]	–
client / citizen-of / general-staff		
social / parent / sibling / spouse / grandparent	[75]	–
associate / other-professional / other-personal		
product-of	[73]	–
inclusion	–	[11,15,19,39]
protein-organism-location	[66]	–
affects / causes / exhibits / analysis /disrupts	[83]	–
dead / wounded / kidnapped / arrested / perpetrator	–	[97]
quantified / modificated / results / indicates	–	[39]
people_of / name_of / character_of / affiliation	–	[15,45]
professional_relation / ownership		
date_of / birth_date / death_date / life_time		
representative_of / represented_by / other_edition		
practised_in / participant_in / has_participant		

the relations presented, we have traditional relations, like *hyponymy* and the corresponding i*nstance-of* relation, which involves named entities. Different relations regarding NEs have been considered by evaluation conferences. For example, for English, ACE proposed the extraction of several NE relation types and subtypes. Conferences focused on relations such as ACE [75] and MUC [73] are references to several proposed research projects for English, Chinese and Arabic. Such a line of research considering Portuguese was proposed in the scope of the Recognition of Relation between Named Entities (ReRelEM) [45] track of HAREM. More details about MUC, ACE and HAREM will be presented in Sect. 5.

It should be noted that there are two areas that deal with relation detection between named entities: Anaphora Resolution and Information Extraction. A brief discussion about these two areas is presented in [44]. Anaphora resolution is the process of identifying expressions which points back to another expression in the text. In these expressions, called anaphoric, the entity to which they refer is their antecedent. The focus of anaphora resolution is determining the antecedent chains, although it also implicitly allows for elicit semantic relations between referents. The identification of antecedent chains of a text can improve the learning process of RE systems. According to Gabbard et al. in [47], sentences within a corpus that contains these chains are used to induce alternative ways of expressing relations. In this paper we focus on the Relation Extraction task between named entities, but not on anaphoric relations between NEs—such as *identity*.

2.2 Relation Extraction

Relation Extraction between NEs is a challenge to IE that seeks to identify instances of pre-defined relations between

Table 2 NER and RE extraction from text

Named Entities	Relations
Steve Ballmer <Person>, is an American businessman who servesas the **CEO** <Employment> of **Microsoft** <Organization>, having held that post since January 2000	**CEO** <employment> **Microsoft** **Steve Ballmer** <employee-of > **Microsoft** **Steve Ballmer** <performs> **CEO**

certain entities. According to the ACE program [30], the goal of RE task is to detect and characterize relations between (pairs of) entities. For example, an RE system has to be able to extract an *employment-organization* relation (defined by ACE) between "*CEO*" and "*Microsoft*" in "*the CEO of Microsoft*" [94].

In order to clarify this, Table 2 shows the NEs and Relation Extraction from plain text described above, where the extracted entities are bold-faced with their corresponding categories.

> "Steve Ballmer, is an American businessman who serves as the CEO of Microsoft, having held that post since January 2000."

For the extraction of explicit relations in texts, as exemplified in Table 2, the analysis of several aspects regarding the syntactic and semantic structures of the sentence is required. Some aspects often analyzed for RE are presented below:

- The occurrence of words that can express a particular relation around or nearby entities. For example, "author of" in "George R. R. Martin is the author of A Song of Ice and Fire".
- Lexical categories which can help identifying whether a word defines a relation or not. For example, the verb "founded", in "Microsoft was founded by Bill Gates", may express a relation of *Affiliation* between the entities "Bill Gates" and "Microsoft";
- The syntactic structures of the sentence that might contain relations, such as prepositional or verbal phrases provided by a parser—for example, the relation of *Location* expressed in "I'm going to work for Microsoft in Washington" between the entities "Microsoft" and "Washington".

Relation Extraction involves the identification of relations between entities already identified in natural language texts. In the literature, we find several works that consider NER to be an integral part of RE systems [1,54,64,93], given that NER can help in the identification of arguments/entities that are part of certain relations. According to [88], the identification of named entities is the first step towards the semantic analysis of a text, being crucial to Relation Extraction systems. However, the NE tagger may benefit with the feedback from subsequent stages in a information extraction pipeline, such as semantic Relation Extraction [59].

Currently, the RE challenge is a matter of research that emcompasses different areas such as NLP, Machine Learning, Databases, Information Retrieval, and others. For the RE task for IE, different approaches were developed: supervised learning techniques employing an annotated corpus; unsupervised approaches based on generic extraction patterns; semi-supervised methods—including bootstrapping that needs only a few annotated examples—and also the Open IE approach for the extraction of relations not previously defined.

In machine learning techniques, patterns for identifying relations are not manually written (handcrafted rules) but are learned from labeled examples [71]. The most commonly applied machine learning methods in RE systems are: Hidden Markov Models (HMM) [40], Conditional Random Fields (CRF) [25,63,65], k-Nearest-Neighbors (KNN) [104], Maximum Entropy Models [61], and Support Vector Machines (SVM) [26,53,98,104]. Such methods will be discussed in Sect. 3. However, the performance of these methods depends on the features that are used. In general, RE systems perform textual analysis and define a set of features to be extracted from the sentences of texts. Textual Analysis involves POS tagging, Parsing, NER and a choice of features guided by intuition and experiments. The most common features used in RE systems are listed below:

- The sequence of words between the two entities;
- Types of entities (*Person*, *Organization*, *Location* etc.);
- POS tag of each word;
- POS tags of the sequence of words between the two entities;
- The number of words separating the two entities;
- The head of the segment;[2]
- The words of the segment;
- A window of K words to the left of *Entity1* and their POS tags;
- A window of K words to the right of *Entity2* and their POS tags;
- Syntactic information of each word;
- Syntactic information of the sequence of words between the two entities.

[2] Segment can be considered as the sequence of words that describes the relation between the two related entities.

2.3 Relation Extraction for Portuguese

Portuguese Relation Extraction has been boosted mainly by the HAREM evaluation contest. The first HAREM dealt with NER mainly whereas the RE task appeared in the Second HAREM in 2008, in the ReRelEM track. A great deal of the literature in this area refers to this evaluation contest. [44] deals with the study of target relations for the ReRelEM track [18] (further details in Sect. 5). Such relations were selected from the manual analysis of twelve Portuguese texts extracted from the Second HAREM's Golden Collection.

According to [45], one of the reasons for studying the corpus for the selection of relevant relations between NEs, is the fact that the literature regarding the analysis of linguistic relations among words or expressions does not often deal with relations between NEs. Besides, the manual analysis of such texts, albeit time consuming, is a valuable source of knowledge representation on a certain language and, in this case, the study of relations between NEs in Portuguese would largely benefit from it.

3 Relation Extraction systems and their approaches

At present, research in Relation Extraction focuses mainly on pattern learning and matching techniques in order to extract relations between pairs of entities from various text sources, such as a collection of news articles, Web pages, and others. However, depending on the application and on the resource available, the Relation Extraction task can be studied for different settings. In this section, we focus on these two cases: in RE systems in which the set of relations of interest can be previously defined (closed RE), or when there is no pre-defined relation type in the input (open RE). A brief discussion of some RE systems and their corresponding approaches are presented in the following sections.

3.1 Closed Relation Extraction

It is usual for the RE system to have the target relation as an input, as well as manually identified extraction patterns, or patterns that were extracted by machine learning through previously tagged relation examples [1,10]. Those inputs are specific for the target relation, thus, the identification of new relations requires identification of new extraction patterns or the definition of new examples for training, both to be executed manually. Given that scenario, the need for manual intervention grows in direct relation to the number of targeted relations [4].

For supervised Relation Extraction, existing work often uses the ACE benchmark data sets for evaluation [26,53,81, 103,104]. A set of relation types is defined and the task is to identify pairs of entities that are related and to classify their relations into one of the pre-defined relation types. For example, Zelenko et al. [103] introduced the kernel methods for RE, and defined the kernel on the constituency parse trees of relation instances. The problem of RE is treated as the problem of pair-of-entities classification: examples consisted of parts of sentence shallow parse trees, where relation roles were identified by tree attributes, such as member or affiliation relation. For example, in the sentence "John Smith is the chief scientist of the Hardcom Corporation", where the entities "John Smith" and "Hardcom Corporation" are mentioned, we can identify a person-affiliation relation between them.

Cullota and Sorensen [26] extended the idea to dependency parse trees, using a slightly more general version of [103] kernels. Qian et al. [81] proposed a new approach to dynamically determine the tree span for tree kernel-based semantic Relation Extraction and achieved a state-of-the-art performance on the ACE 2004 benchmark data set (the best F-measure score for the seven major relation types).

Recently, semi-supervised and bootstrapping approaches have gained special attention. Bootstrapping-based Relation Extraction systems process large corpus or work on large amounts of data from the Web efficiently, requiring little human intervention [1,10,33,34,79,94,105].

Bootstrapping approaches start with a small number of seed examples to train an initial model. This model is used to label some of the unlabeled data. Then, the model is retrained using the original seed examples and the self-labeled examples. This process iterates, gradually expanding the amount of labeled data. This approach has the advantage of not needing large tagged corpora. For example, the DIPRE system—Dual Iterative Pattern Relation Expansion—which aims at finding book citation patterns from the Web [10], uses bootstrapping technique. *Author-title* relations are extracted from 24 million Web pages starting from an initial set of 5 books. The SNOWBALL system [1] extracts *Organization–Headquarters* relation from Web pages and includes the use of named-entity tags based on the annotation of a POS-tagger from the Alembic Workbench [29].

The bootstrapping approach for exploring generic patterns is also proposed. Espresso is a weakly-supervised general-purpose system, which is designed to extract various semantic relations from texts [79]. KnowItAll is an unsupervised, domain-independent system that extracts facts from the Web [33,34], which has the particularity of using a novel form of bootstrapping that does not require any manually tagged training sentences. This latter system starts with a domain-independent set of generic extraction patterns from which it induces a set of seed instances. Unsupervised information extraction does not require hand-tagged training data; then, unsupervised extraction systems can recursively discover new attributes, instances and relations automatically without human intervention.

In general, due to the many iterations, bootstrapping-based extraction systems suffer from semantic drift [28,69], which occurs when errors in labeling accumulate and the learned concepts drift from what was intended. Carlson et al. [17] proposed that this problem could be dealt with by coupling the semi-supervised learning of categories and relations. The authors describe the use of the CBL algorithm (Coupled Bootstrap Learner), which takes as input an initial ontology and a large corpus from Web pages. Motivated by the fact that these systems return a significant number of errors or low relevance relation instances, a graph-based method to rank the returned relation instances of bootstrapping Relation Extraction system is proposed in [64]. Ordering the extracted examples by the relevance to the given seeds is helpful to filter out irrelevant instances.

Besides bootstrapping, another promising weakly supervisioned approach, called distant supervision, has been applied for Relation Extraction. According to Mintz et al. [70], if two entities participate in a relation, any sentence that contains these entities also expresses this particular relation. Mintz et al., make use of features extracted from different sentences containing pairs of entities from the knowledge base Freebase [9] to build a rich vector of features. Such features are based in lexical information, syntactic and NER. In [58], the MultiR system for multi-instances learning is presented, which also uses a distant supervision approach from Freebase and applies to the proposed features in [70].

Recently, the literature has presented the probabilistic model CRF as a good alternative, in which RE is handled as a text tagging task [25]. The CRF model has been applied efficiently in many tasks of sequential text processing [65], standing out in different applications for RE. Bellare and McCallum [5] extract 12 biographic relations by applying a CRF extractor, which is trained from BibTeX research records from article citations. In [21] CRF is applied to extract relations between knowledge elements, involving the relation kinds: *Preorder*, *Illustration* and *Analogy*. Sahay et al. have used CRF to identify relations from biomedical abstracts, considering the triple (Concept1, Relation, Concept2) [83].

Culotta et al. [25] propose the integration of supervised machine learning that learns relational and contextual patterns for the extraction of a familiar relationship from biographical texts. For that, it is proposed an extraction model of relations using CRF, which verifies whether the entities found in biographic texts are related to the topic of the page, starting from a set of relations previously known. Li et al. [65] apply the CRF model for extraction of specific relations between two entities based on general relations. Take, for instance, the "Employment" relation (job title/position a person holds at the organization), one of the major relation types defined in ACE. Depending on the objective of the Relation Extraction task, the exact information about the job may be needed.

3.2 Open Relation Extraction

In the literature, there are methods that do not need previously tagged corpora or an initial set of tagged examples, nor do they aim for the extraction of predefined relations. An example is the Finite State Automaton Text Understanding (FASTUS) system, which is based on a Finite-State approach [57], and also on methods that are completely unsupervised, such as the one proposed by Hasegawa et al. [54]. The FASTUS system works as a cascaded, non-deterministic finite-state automaton: it separates processing into stages, and each level corresponds to a linguistic natural kind. The authors believe that a decomposition of the natural language problem into levels is essential to the approach. The Relation Extraction approach presented in [54] is based on a corpus defined by clustering NE pairs according to its context similarity (cosine similarity) and the usage of the NE tagger [54] for NE identification. The authors assumed that NE pairs that appear on a similar context may be grouped and that each is an instance of the same relation.

An approach for RE, which does not need a pre-specified definition of relation, was proposed by Banko et al. [3]. This approach is called Open Information Extraction (Open IE), and it is currently being studied by few research groups. Open IE approach targets larger corpora such as the World Wide Web, that contains an expressive number of relations that may not be previously recognized and explored. An Open IE system aims at extracting a large set of related triples (E1, Rel, E2) from a certain corpus without requiring human supervision, where E1 and E2 are strings meant to denote entities or noun phrases, and Rel is a string meant to denote a relationship between E1 and E2 [2].

The sole input for an Open IE system is a corpus, and relations are extracted by applying a set of heuristics and domain-independent patterns. One generic pattern of extraction that stands out is noun phrases participating in "subject-verb-object" relationship [4,48]. For example, from the sentence "Einstein received the Nobel Prize in 1921", we can extract the triple (Einstein, received, the Nobel Prize). Thus, it is clear that an Open IE system has to deal with a substantial amount of challenges that traditional RE systems do not, as Banko and Etzione [4] presented:

1. A traditional RE system usually searches for entities that are associated with the type of relation the system was configured to extract, given the nature of the relation, which is already known. An Open IE system tries to find evidence of existing relations as well as the entities taking part in those relations;
2. A traditional RE system searches for specific patterns for each relation. Open IE systems need a set of patterns that are not related to any specific relation, and these features must be useful to extract relations of any nature;

3. A traditional RE system usually employs different types of NEs to help in the process of identifying entities that are comprehended by a particular relation, e.g., the *Located* relation has as arguments the *Location* and *Organization* entities. In Open IE the relations are not predefined, thus, the types of arguments/entities are also not known.

DIPRE, SNOWBALL, Espresso and KnowItAll systems presented previously are all relation-specific systems. The first Open IE system was the TextRunner system [3,102], which used a Naive Bayes classifier with POS and NP-chunk features. This system uses a small set of hand-written rules to heuristically label training examples from the Penn Treebank. Banko and Etzioni [4] present the O-CRF system based in a Conditional Random Field model (CRF). The authors show that many relations can be categorized using a compact set of lexicon-syntactic patterns.

There are seed-based systems that can perform both traditional RE and Open IE—for example, the StatSnowball system [105], which extends the SNOWBALL system with the addition of the use of statistical methods for extracting relations between entities, specifically the Markov Logic Network.

A new approach to Open IE which uses Wikipedia as a source of training data is proposed in [99,100]. The authors present WOE system—Wikipedia-based Open Extractor—which generates relation-specific training examples by matching between Wikipedia Infobox content with corresponding patterns. In contrast with TextRunner and StatSnowball systems, which employ only shallow features in the extraction process, WOE can learn two kinds of extractor: WOE$_{\text{PARSE}}$ learned from dependency path patterns; and WOE$_{\text{POS}}$ trained with shallow features like POS tags.

The recent work described in [36] shows that the output of Open IE systems (such as TextRunner and WOE) is abundant with uninformative and incoherent extractions. To treat these problems, the authors implemented the syntactic and lexical constraints in the ReVerb Open IE system. A good example is the occurrence of the pattern: a verb followed immediately by a preposition (e.g., *located in*). These constraints serve two purposes: (i) to eliminate incoherent extractions, and (ii) to reduce uninformative extractions by capturing relations phrases expressed by a Verb-Noun combination, such as light verb constructions (LVCs).

A multilingual dependency-based OIE system (DepOE) has been recently proposed in [48]. DepOE system follows three steps: (i) dependency parsing, where each sentence is analyzed using multilingual parser, called DepPattern[3]; (ii) clause constituents, where for each parsed sentence the verb clause is found, and their participants (such as subject, direct object, other), and (iii) a set of rules is applied on the clause constituents in order to extract the target triples. This system was used to extract triples from the Wikipedia in four languages: Portuguese, Spanish, Galician and English. It should be stressed that, for English, pos-tagging information based on Tree-Tagger[4] were used and for other languages the information was based on the annotations of FreeLing.[5]

In general, previously mentioned works in Open IE have focused mainly on syntactic features for relations extraction. Christensen et al. [22,23] investigate the use of semantic features for the task of Open IE, specifically the application of Semantic Role Labeling (SRL). SRL consists of detecting semantic arguments associated with a verb in a sentence and their roles, such as *Agent*, *Patient*, and others. The authors study the trade-offs between TextRunner, a state-of-the-art Open IE, and SRL-based extractor across a broad range of metrics and experimental conditions, both qualitative (such as N-ary and binary relations) and quantitative (such as small and large corpus).

3.3 Systems for Portuguese

As stated before, many approaches have been used for Relation Extraction in English, which have been presented in the previous sections. In contrast, there are very few proposals for Relation Extraction in Portuguese. One of the main obstacles for the progress in this domain within that research field is the lack of resources such as annotated data. There is also a demand for the development of new techniques, tools and specific resources such as lexical bases, domain ontologies.

In this section we present the approaches used by systems that took part in the ReRelEM track [11,15,19] and also papers approaching Relation Extraction in Portuguese that are available in the reviewed literature [16,38,39,42,96,97,101]. It is worth highlighting that for most systems the set of relations were previously defined (closed RE). The only system which applies Open IE approach for Portuguese language is DepOE, already presented in the Sect. 3.2.

Relation Extraction for Portuguese systems is usually rule-based [11,15,19]. According to [89], many extraction tasks can be executed by employing a set of rules, which might be coded manually (hand-coded) or learnt through examples. These systems applied simple heuristics that explore evidences of relations between NEs in the texts, comprehending different analysis: lexical, syntactic and semantic analysis, entity types and information from external sources. From the external sources it is worth highlighting the Portuguese Wikipedia that provides a great number of structured information as well as ontologies which provide names.

[3] http://gramatica.usc.es/pln/tools/deppattern.html.

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[5] http://nlp.lsi.upc.edu/freeling/.

The REMBRANDT[6] system [15] was developed to recognize all kinds of NEs and relationship between them. This system makes use of Portuguese Wikipedia as an external resource, as well as some grammar rules that describe internal and external evidence about NEs. According to the author, using Wikipedia solved the classification and disambiguation of NEs that are part of the recognized relations, since an NE can have more than one meaning. Grammar rules represent patterns found in sentences that function as a signal for the presence of an NE with specific semantic properties. The extraction of relations between NEs uses basic heuristics considering their constituents, categories and connections among their respective pages on Wikipedia.

The recent work described in [16] shows that the REMBRANDT is now a mature tool and it can therefore be used by the NLP community on several IE tasks. The REMBRANDT is composed of the REMBRANDT NER tool (participated on the Second HAREM); RE-NOIR, a semantic query reformulation module for document retrieval; SASKIA, a knowledge base for all knowledge resources and stored data; an indexer that generates standard term and semantic indexes for all extracted NEs; and a retrieval and ranking module, called LGTE[7] (Lucene with GeoTemporal Extensions).

The SeRELeP[8] system [11] aimed at recognizing three relations: *Identity*, *Inclusion* and *Location*. The steps for identification/classification of NEs were carried out by PALAVRAS parser [6].

SEI-Geo [19], in contrast with the other systems, is focused on ontology enrichment [20]. SEI-Geo is an extraction system that deals with NER concerning only the *Location* category and its relations. This system makes use of Geo-ontologies, making the search of relations between known locations on pieces of text based on the relations within the ontology possible. According to [98], semantic relations provide richer metadata connecting documents to ontologies and enable more sophisticated semantic search and knowledge access.

As SEI-Geo, systems that focus on Relation Extraction from text can help finding ontology instances. In the literature, we found few studies that address ontology population for Portuguese. Xavier and Lima [101] present a semiautomatic method to extract and populate domain ontologies from the category structure of Portuguese Wikipedia. The instantiation task is performed on the same stage in which the relations between concepts are extracted.

In [39] a system for information extraction from medical reports was presented. The authors reported a golden collection in the scope of the MedAlert project, in which clinical documents relative to hospitalization episodes are annotated with its multiple entities and relations. The entities of interest of MedAlert were defined as something real referred in the text, for example, the mentioned medication, the exams performed etc. Yet, the relations are the connections between these entities, such as the results of an exam or the medication indicated for a pathology. For automatic extraction of entities and relations, the REMMA[9] system (Reconhecimento de Entidades Mencionadas do MedAlert) was used. REMMA is capable of using several types of resources, whether specialized almanacs (for example, a list with the names of the most common clinical problems), or semantic categories extracted from the analysis of the first phrase of a Wikipedia article.

In [85] a system that identifies family relationships is presented. Historical and biographic documents are texts that are rich in that kind of relation. In the HAREM conference, the *Family* category was a subcategory of *Other* [45], but specific kinds of family relationship such as "*father*", "*mother*" etc., have not been considered. The authors have considered familiar relationship using rule based approach. Among the set of features utilized, genre and number of the words, POS, and syntactic structures such as appositive, head of the sentence, and others stand out.

In [97] a multilingual methodology for adapting an event extraction system to new languages was described. Generally, the task of event extraction is to automatically identify events in free text. This task relies on identifying named entities and the relations between them. The authors created the system NEXUS, which is part of the Europe Media Monitor Family of Applications (EMM).[10] This system aims at identifying violent events, human made and natural disasters, and humanitarian crises in the News reports. Currently, NEXUS can handle four languages (English, French, Italian and Russian). In this work, the authors decided to adapt NEXUS to Portuguese and Spanish languages so as to extend the coverage of this system to Latin American and African areas as well.

There are works for Portuguese that investigate similar tasks such as the extraction of quotes, because it combines several different views over all new topics, for example: named entities, relations extraction and current topics. Quotation Extraction consists of identifying quotations and their author from a text [90]. Fernandes et al. [38] presents the first Quotation Extraction system that uses a machine learning approach for Portuguese (Entropy Guided Transformation Learning supervised algorithm). The focus of this work is on the *quotation-author* relations and it involves two subtasks:

[6] Recognition of Named Entities Based on Relations and Detailed Text Analysis.

[7] http://lucene.apache.org/.

[8] System for Recognition of Relations for the Portuguese language.

[9] The REMMA system was at first developed aiming the participation in the Second HAREM.

[10] Overview article on the EMM family of applications available at http://emm.newsbrief.eu/overview.html.

quotation identification and association between quotation and author.

As presented here, most of the RE systems for Portuguese are based on heuristics and few external resources, such as Wikipedia and domain ontology (e.g., Geo-ontologies), they usually do not make use of machine learning techniques, contrary to the situation for English. Although Portuguese has a large number of speakers, the scientific community dealing with the computational processing of this language is not very large, the Portuguese processing meeting PROPOR (International Conference on Computational Processing of the Portuguese Language)[11] has had around 100 participants in the last years, for example. Linguistic resources for RE research on both English and Portuguese texts are presented in the following session.

4 Linguistic resources for research in Relation Extraction

For further advances in RE research, the availability of comprehensive resources such as dictionaries, lexical bases, domain ontologies and others are of utmost importance. Those allow the development of solutions for specific needs, such as business analysis [64]. Knowledge bases that cover common sense human knowledge is a goal dreamed for a long time in Artificial Intelligence.

Lexical knowledge bases with wide coverage have an increasingly important function for the development of computational tools that attempt to interpret information expressed through natural language.

For English language, Princeton WordNet (WordNet.Pr)[12] [37] is widely used as a resource for Relation Extraction [26,53,56,71,93,98]. WordNet.Pr is a resource manually created by specialists and based on synsets, groups of synonyms expressing a particular concept of natural language. WordNet.Pr has currently a total of 117.000 synsets available, and each synset has a definition, not too dissimilar from the ones found in a dictionary, as well as semantic relations that might occur between synsets, such as hyperonymy, meronymy and others. According to [49], it is unquestionable that the existence of a WordNet accelerates the development of RE. This utility alone is a great motivator for the expansion of this lexical resource towards other languages. Still, the development of resources such as the WordNet is a time-consuming task that requires considerable amount of manual work.

Another important resource for RE systems is the availability of syntactic treebanks. A treebank is a corpus containing sentences with annotations regarding their syntactic structure. For that reason, treebanks can be used to train and evaluate syntactic analyzers. The use of a syntactic analyzer of good quality impacts on the performance of RE task. That is due to the fact that many RE systems depend on natural language processing tools as a pre-processing step, which are error prone and can hinder the performance of the system [2].

For the English language, the most known treebank is Penn Treebank[13] [67], that is a large annotated corpus consisting of over 4.5 million words. Recently, the WOE system [100] was applied in WSJ from Penn Treebank corpora.

Recent RE works are using the encyclopedia Wikipedia as a resource/data source for its applications [15,36,100,101]. Currently, Wikipedia features 22 million articles, which were written collaboratively by volunteers around the world. From these articles, we can count around 3,900 million articles for the English language. Those articles may contain, other than text, information such as tables, images, references to other articles contained within Wikipedia as well as references to external pages. Also, Wikipedia uses a template called Infobox,[14] which is a table featuring basic information about the entity/subject described throughout the article. In the literature there are RE works for English that use information contained within Infoboxes as a data source for training extraction systems [100].

4.1 Linguistic Resources for Portuguese

According to [87], there is a significant amount of material and resources regarding relations between words in Portuguese. For example, there is the WordNet.Br project, mainly focused on Brazilian Portuguese[15] [92], which is an initiative that started in 2003 and currently has around 41.000 word-forms and 18.200 synsets.

The first version made available of the WordNet.Br (Base de Verbos) [16] contains 5,860 verbs in 3,713 synsets, and it was built in alignment to the 2.0 version of WordNet.Pr. WordNet.Br used information contained in the lexical base TeP[17] (Thesaurus Eletrnico do Portugus), which aimed to be a thesaurus integrated with a text processor [91]. TeP follows a data representation scheme identical to WordNet and its second version—featuring a Web interface—was made available for use without charge in 2008; it contains 44,296 terms and 19,885 synsets.

[11] http://www.propor2012.org/index.html.

[12] http://wordnet.princeton.edu/wordnet/.

[13] http://www.cis.upenn.edu/treebank/.

[14] http://en.wikipedia.org/wiki/Category:Infobox_templates.

[15] http://www.nilc.icmc.usp.br/carol/wnbr/wn.html.

[16] http://caravelas.icmc.usp.br/wordnetbr/.

[17] http://www.nilc.icmc.usp.br/tep2/.

For European Portuguese there are two WordNets: WordNet.PT[18] [68]—a manually developed database of linguistic knowledge; its current version has around 19,000 expressions—and the MultiWordNet of Portuguese (MWN.PT)[19]—a lexical semantic network shaped under the ontological model of WordNets. MWN.PT (version 1) spans over 21,000 words and 17,200 manually validated synsets.

Recently the lexical ontology for Portuguese Onto.PT [20] [49,50] was made available. It was created automatically through the exploration of lexical resources for Portuguese and its structure is similar to a WordNet. The current version of Onto.PT (February 2012) features around 160,000 lexical forms and 110,000 synsets. Among the lexical resources explored during the construction of Onto.PT is PAPEL—Palavras Associadas Porto Editora-Linguateca[21] [76]—an important lexical base created by Linguateca.

PAPEL consists in a set of relations established among terms, which are semi-automatically extracted based on the dictionary Dicionário PRO da Língua Portuguesa. The built process of PAPEL is based on a set of rules that makes use of certain lexical-syntactic patterns to extract semantic information between the meaning of words that occur in a definition (p) and the defined meaning of the word (v) in the format of triples (p relation v). The current version of PAPEL (v3.2) has about 190,000 relations (or triples). It is worth noticing that the relations which integrate PAPEL have been chosen from the analysis of the dictionary content and in the review of literature about relations between words and ways of structuring dictionaries. Among the relations we can mention: *synonymy*, *hyperonymy*, *part*, *member*, *causer*, *producer* and *local*.

Currently, it was made available as a service for validation of relations called VARRA [46], which is developed with the main objective of assisting the manual evaluation/validation of semantic relations in Portuguese. According to the authors, the participant words of a semantic relation are always regarded in authentic context, specifically represented by sentences from corpora available in the AC/DC project,[22] so that the realization of the validation between the pairs of words is possibly the most similar to human interpretation. The authors have presented the initial results of the semantic relations, which integrate the PAPEL.

For the Portuguese language, another evaluation resource is Floresta Sintática[23] [43], which is a publicly available treebank for Portuguese. It can be used within NLP research as a tool to test/train and evaluate a syntactic analyzer [8], which can benefit systems that need morphosyntactic annotations, such RE systems [11,38,42,96].

Texts on Floresta Sintática come from two distinct corpora: CETEM Público,[24] composed by news content written in European Portuguese and CETENFolha,[25] composed by news content written in Brazilian Portuguese, both automatically annotated by the syntactic analyzer Palavras [6]. The resources produced by the Floresta Sintática project also comprehend Floresta Virgem, its unrevised syntactic tree set, and Bosque, which corresponds to the revised part of Floresta Sintática, composed by 9.368 sentences and 190.513 lexical items.

In [8], it is mentioned that Floresta Sintática is an important resource of joint evaluations for Portuguese, considering the relevance of information contained within it. It is worth mentioning that Floresta Sintática was used for building the Golden Collection of the joint evaluation initiative HAREM [88], and also the specific use of Bosque in the international joint evaluation initiative CoNLL-X, that happened in 2006 [12].

As we mentioned in the Sect. 3.3, an important available knowledge source for the Portuguese language is Wikipedia, currently featuring around 738.000 articles for the Portuguese language (as in June 13th 2012). The REMBRANDT system [15], for instance, utilizes Wikipedia as an external resource for the NE classification step.

Despite the resources previously presented in this section, Portuguese is still far from having well documented material and following a consensus about the various semantic relations comprehended by the language lexicon, which harms greatly initiatives of automatic processing. As mentioned before, the WordNets for Portuguese featured a small number of synsets when compared to WordNet.Pr for English. Onto.Pt was made available just recently and it is so far being constructed, still it is an important lexical-semantic resource for Portuguese [51].

Wikipedia can be seen also as an important resource for RE tasks, but it is also important stating that the Portuguese Wikipedia[26] features a set of articles that amount to around one fifth of the English Wikipedia, which might limit the usage or comparison of this resource.

Generally, there are far more resources available for English than for Portuguese, limiting research within that field for this specific language.

This fact is better illustrated by Table 3, which displays the number of external tools employed by RE systems for each of these languages.

[18] http://www.clul.ul.pt/clg/wordnetpt/index.html.

[19] http://mwnpt.di.fc.ul.pt/index.html.

[20] http://ontopt.dei.uc.pt/.

[21] http://www.linguateca.pt/PAPEL/.

[22] The AC/DC project is an interface to access and availability of corpora in Portuguese, which contains 22 corpus. Available at: http://www.linguateca.pt/ACDC/.

[23] http://www.linguateca.pt/floresta/.

[24] http://www.linguateca.pt/cetempublico/.

[25] http://www.linguateca.pt/cetenfolha/index_info.html.

[26] http://pt.wikipedia.org/wiki/Wikipdia_em_portugus.

Table 3 External NLP tools and resources used in related work

External NLP tools and resources	English	Portuguese
POS tagger	[1,33,34,48,54,79]	[38,39,48]
Parser	[48,100]	[11,42,48,96]
WordNet	[26,53,56,93,98]	–
Wikipedia	[36,48,100]	[15,39,48,101]
Ontology	[93]	[19]
NP identifier [84]	–	[42,96]
NER system	[1,54,64,93]	[11,15,19]
OpenNLP [77]	[3,4,17,36,100]	–
GATE [27]	[93]	–

The lack of resources for RE system evaluation as much as annotated data is a further obstacle to the advancement of research in this area. One way is the realization of evaluation conferences, which provide resources such as baselines, reference corpora and program evaluation results. The evaluation conferences are presented in the next section.

5 Evaluation

One important prerequisite for the evaluation of NLP applications is to know extensively about the proposed problem, so the development of evaluation methodologies will encompass the proper quantification and qualification of relevant results. One can only develop a good evaluation methodology if the analyzed problem is properly quantified and the possible advance of the proposed approaches are identified.

Overall, the performance evaluation metrics used to evaluate NER and RE are the same as the ones for Information Retrieval [32]. The most commonly used measures for such evaluations are Precision, Recall and F-measure, defined as follows:

$$\text{Precision} = \frac{\text{Number correct}}{\text{Number correct} + \text{number incorrect}} \quad (1)$$

$$\text{Recall} = \frac{\text{Number correct}}{\text{total Number of relations to be found}} \quad (2)$$

$$F\text{-measure} = \frac{2 * P * R}{P + R} \quad (3)$$

However, the evaluation of the RE task depends on reference corpora and/or datasets, which work as a comparison element for analyzing and evaluating systems dealing with this task. Reference corpora is needed for providing a norm with which frequencies of the studied corpus will be compared. Such corpus is named a Golden Standard and contains annotations—usually manually made by more than one specialist—following defined guidelines that describe how the annotation must take place as well as approaches to reach consensus among specialists for this specific task. An example of a Golden Standard are the MUC's datasets [72,73], which are used to evaluate the Named Entity Extraction and Relation Extraction task.

Evaluation Contest conferences as MUC, ACE, TAC and HAREM compile the participation of several systems that are compared while executing NER and RE tasks in [86]. The goal of a joint evaluation is to improve the state-of-the-art within the field, since it promotes research in the area related to the proposed task; as a result, it produces evaluation methodologies and resources such as test databases. Conferences focused on the evaluation of intelligent systems that approach different tasks related to the comprehension of language have helped the advance of NLP studies [14]. In what follows, we present a brief description of joint evaluation conferences dealing with RE.

The first important conference to define what the NER task comprehended as well as its initial evaluation models was Message Understanding Conference (MUC). Its first edition took place in 1987 and aimed at developing a joint evaluation of IE. MUC's seventh edition has fostered one extra task regarding identification of relations among categories (Template Relation, TR) [73]. This task comprises the extraction of well-defined aspects of text from newspapers written in English, the TR task relations were shown in Table 1 (see Sect. 2.1).

Other evaluation initiatives that were also quite important are the Automatic Content Extraction (ACE) program and the evaluation session proposed by the Text Analysis Conference (TAC). The first edition of the ACE program was held in 1999 and it carried out a pilot study for English language. From 2000–2001 onwards, ACE has expanded the definition and scope of the NER task for English and Chinese language (Entity Detection and Tracking, EDT). The ACE evaluation held in 2002–2003 included Relation Detection and Characterization (RDC) [30] and this task was carried out by 2008 [75]. RDC comprises identifying/classifying types of relations and corresponding subtypes between entity pairs. Table 1 in Sect. 2.1 presents some types and subtypes from ACE relations.

Following MUC and ACE, TAC (Text Analysis Conference) began in 2008 and is held annually since then. TAC is a series of evaluation workshops organized to encourage research in NLP and related applications, by providing a large test collection, common evaluation procedures, and a forum for organizations to share their results. Currently, TAC 2012[27] focuses on a KBP track involving three areas (Entity-Linking, Slot-Filling and Cold Start Knowledge Base Population), all aimed at improving the ability of automatically

[27] http://www.nist.gov/tac/2012/KBP/index.html.

populating knowledge bases from text, including English, Chinese, and Spanish languages.

Besides the evaluations performed in these conferences, many research works consider different datasets. In supervised systems, the RE task is expressed as a classification one [71], therefore, measures (such as Precision, Recall, F-measure) can be used to evaluate those systems. Reference data can be used for learning and to calculate those measures by cross-validation. As such, RE systems that employ unsupervised methods need a reference corpus annotated with expected information for their validation, or the validation can be achieved through the manual evaluation of the extracted relations. For example, [54] evaluates the relations detected automatically using clustering methods, the authors analyzed the dataset manually.

In a similar fashion, for the application of semi-supervised methods, there is hardly one tagged test set for the validation of the learned model. Besides, semi-supervised methods for Relation Extraction are usually applied over a great volume of data, such as Web pages, and the outcome is usually a large number of new relation patterns (e.g., Open IE). Therefore the manual analysis of such results would be very time consuming. What is usually done is the manual analysis of a subset of the data. This subset may be randomly extracted or focused on a specific group of relations selected from the whole set. For example, the DIRPE system obtained as a result a list over 15,000 books; the authors randomly selected 20 books and analyzed them manually. The TextRunner system was executed in a large corpus comprised by 9 million Web pages. The relations were randomly selected, totalizing 10 relations identified in at least 1,000 sentences in the corpus. In the evaluation, 400 tuples were analyzed manually by three human experts. Both systems are described in Sect. 3. Table 4 shows an overview of the evaluation and different datasets used by some of the related work for English.

5.1 Evaluation of Portuguese

For Portuguese, similar efforts of joint evaluation took place. HAREM (Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas)[28] is a contest exclusively dedicated for the Portuguese language, and it has studied expressions regarding proper names recently.

Among the main features of HAREM, the semantic model stands out, which for the semantic classification of named entities requires their meaning in the context, instead of the meaning expressed in dictionary. Consider, for instance, the following sentence: "O Vasco da Gama passou enfim para a primeira divisão." ("Vasco da Gama has finally made it to the first league."). Hence, "Vasco da Gama" can be classified at least as *Person* category or *Organization* category, although by means of discourse context we can relate "Vasco da Gama" with a sport club. This shows that the HAREM task is considerably more difficult and fine-grained than the classical NER task [41], as performed for example in MUC.

The first event for HAREM evaluation happened in 2005 and it has followed MUC evaluation criteria, but since then the process has gone through alterations. HAREM is a milestone in joint evaluation efforts focused on Portuguese, given that, prior to HAREM, only two research works dealing the NER evaluation for Portuguese were found [7,78]. The Second HAREM took place in 2008 and it allowed systems to choose categories, types/subtypes, and it also included the task of identifying the semantic relations between NEs—ReRelEM track—it was concerned with the automatic detection of relations between NEs in a document [41].

Among these resources, there are the Golden Collection[29] with relations manually annotated; SAHARA[30] (Automatic HAREM Evaluation Service) online tool that makes possible quick evaluation of outputs for systems; Etiquet(H)AREM,[31] which helps linguists to annotate and compare annotations, and others [41].

The relations defined in ReRelEM are *Identity* (or *coreference*—entities with the same referent, defined to all the categories and whose instances must have the same category), *Inclusion* (*Included/Includes*—defined to all the categories, and whose instances must have the same category), *Location* (*Occurs_in/Located_in*—defined between an *Event* and a *Location* or between an *Organization* and a *Location*, respectively) and *Other* (relations that do not correspond to any other previously listed category). The latter type grouped a set of 22 new relations. Table 1 includes those relations highlighted in italic instead of *Other*.

As a result of the annotation, we have the ReRelEM's Golden Collection with 6,790 relations distributed as follows: *Identity* (*Identidade*)—1,436; *Inclusion* (*Inclusão*)—1,612; *Location* (*Localização*)—1,232; and *Other* (*Outra*)—2,510 [18,44]. After the ReRelEM defining and presenting the results, Linguateca extended/revised the annotation of relations to the whole Second HAREM Golden Collection (LÂMPADA 2.0).[32] We can mention some of the changes related to the directives of ReRelEM utilized in Second HAREM and documented in [18]: the inclusion of the relation *practitioner_of* (*praticante_de*)/*practiced_by* (*praticado_por*) and the relation *death_in* (*local_morte*), and the elimination of the *representative_of* (*representante_de*) /*represented_by* (*representado_por*) relations.

[28] Evaluation of Systems for Named Entity Mention.

[29] http://www.linguateca.pt/HAREM/coleccoes/CDSegundoHAREMReRelEM.xml.

[30] http://www.linguateca.pt/HAREM/avaliador.

[31] http://linguateca.dei.uc.pt/harem/Manual_etiquetharem.pdf.

[32] Further details can be found in http://www.linguateca.pt/HAREM/.

Table 4 Data and evaluation methods for English

References	Data/corpora	Data size	Method	Evaluation	Performance (%)
Brin [10]	Web pages	24 million pages	Exact pattern matching	Manual evaluation of 20 books selected from a list of over 150,000	19 correct books—95 %
Agichtein and Gravano [1]	North American News corpus	300,000 newspapers	Matching with similar function	Manual evaluation of a set of 100 tuples	93 correct tuples—93 %
Hasegawa et al. [54]	Articles from New York Times	1 year (1995)	Clustering	Manual evaluation of the relations for 2 domains	Person-GPE $F = 80$ %, Company-Company $F = 75$ %
Pantel and Pennacchiotti [79]	Articles from TREC-9 and CHEM	TREC-9 = 5,951,432 words, CHEM = 313,590 words	Weakly-supervisioned classifier	Manual annotation of 680 instances from TREC and CHEM corpora (2 experts)	TREC part-of $P = 69.9$ %, succession $P = 49$ %, CHEM is-a $P = 76$ %, reaction $P = 91.4$ %, production $P = 55.8$ %
Carlson et al. [17]	Web pages	200 million pages	Coupling Semi-supervised Learning	Freebase database as Golden Standard	Category average $P = 83$ %; relation average $P = 84$ %.
Li et al. [64]	Wikipedia and Tago project	Wikipedia = 4,556,821 pages, % Tago = 67,973 entity pairs	Semi-supervised multi-view ranking	5 types of relation extract by YAGO Project as Golden Standard	Relation average = 39 %
Banko and Cafarella [3], Yates et al. [102]	Web pages	9 million pages	Naive Bayes	Manual evaluation of 400 tuples (3 experts)	80.4 % correct tuples
Banko and Etzioni[4]	Sent500 corpus [13]	Sent500 = 500 sentences	Conditional Random Fields	Small set of labeled data for 4 relations from Sent500	Open relation $F = 59.8$ %; pre-specified relation $F = 29.5$%
Zhu et al. [105]	Sent500 corpus and Web1M corpus	Sent500 = 500 sentences, Web1M = 1 million of blocks of Web pages	Markov Logic Networks	Manual evaluation of the extracted tuples from Sent500	$F = 76.4$ %
Wu [99], Wu and Weld [100]	WSJ from Penn Treebank, Wikipedia and Web pages	–	Conditional Random Fields	Manual evaluation of 300 sentences from each corpora (2 experts)	WSJ $F = 64.7$ %, Wikipedia $F = 57.2$ %, Web $F = 65$%
Culotta et al. [25]	Articles from Wikipedia	271 articles	Conditional Random Fields	Manual annotation of the 53 family relations	$F = 61.36$ %
Li et al. [65]	Articles from New York Times, articles from Wikipedia [25]	New York Times = 150 articles, Wikipedia = 271 articles	Conditional Random Fields	Manual annotation of the relations	New York Times Employment $F = 80$ %, Wikipedia personal/social $F = 51$ %
Fader et al. [36]	Web pages	500 sentences	Logic Regression classifier	Manual evaluation of each extraction as correct or incorrect (2 experts)	$F = 69.8$ %.
Liu et al. [66]	Expert-curated corpus	150K words	Semantic interpretation approach	Manual annotation of 565 relation instances for protein-organism-location	$F = 74.9$ %
Gamallo et al. [48]	Sentences from Wikipedia in English, Spanish, Galician, Portuguese (2010)	English = 78,826,696, Spanish = 21,208,089, Galician = 1,461,705, Portuguese = 11,714,672	Unsupervised extraction of verb-based triples	Manual evaluation of 200 sentences from English Wikipedia (2 experts)	$P = 68$ %

Table 5 Data and evaluation methods for Portuguese

References	Data/corpora	Data size	Method	Evaluation	Performance (%)
Brucksen et al. [11]	HAREM/ ReRelEM Golden Collection	4,417 words	Set of heuristics based on morphosyntactic and semantic information	Golden Standard annotated manually	All relations $F = 36$ %
Cardoso [15]	HAREM/ ReRelEM Golden Collection	4,417 words	Set of grammar rules	Golden Standard annotated manually	All relations $F = 45$ %
Chaves [19]	HAREM/ ReRelEM Golden Collection	4,417 words	Set of grammar rules	Golden Standard annotated manually	All relations $F = 27$ %
Xavier and de Lima [101]	Tourism category from Wikipedia	–	Semi-automatic method based on structure from Wikipedia and syntactic heuristics	Golden Standard the domain of Tourism	$F = 85$ %
Santos et al. [85]	Biographies texts from Wikipedia, CETEMPblico corpus	CETEMPblico = 110 sentences	Rule-base approach	Manual evaluation of the family relations	Wikipedia $F = 29$ % CETEMPblico $F = 36$ %
Ferreira et al. [39]	MedAlert corpus	2,724,860 tokens	REMMA system	MedAlert Golden Standard composed by 20 texts annotated manually	Inclusion $F = 89$ %
Tanev et al. [97]	News articles for Portuguese about security and disaster-related topics	News articles = 3.4 million titles, disaster-related articles = 100 (April 2009)	Ontopopulis system	Comparative evaluation between Baseline Portuguese and the results	Dead $F = 69$ %, Wounded $F = 51$ %, Kidnapeed $F = 67$ %, Arrested $F = 47$ %
Fernandes et al. [38]	GLOBO QUOTES from Globo.com	Around 13.5 million tokens	Entropy Guided Transformation Learning	Baseline system manually constructed	Quotation-Author $F = 79.02$ %

Systems that take part on joint evaluation conferences for Portuguese, such as HAREM, follow the conference directives. For example, the REMBRANDT, SEI-Geo and SeRELeP systems used ReRelEM's Golden Collection during the evaluation of the ReRelEM track. In general, the relations annotated by these systems were compared with the ones in the Golden Collection, and each triple (NE Relation NE) was scored as correct, missing or incorrect [41]. As results of the ReRelEM track, REMBRANDT system achieved the best results in the global scenario (considering all relations: Identity, Inclusion, Location and Other), SEI-Geo got best scores for the Inclusion relation, and SeRELeP reached best results for the Location relation.

There are Relation Extraction works for Portuguese that also demand manual evaluation of the automatically extracted relations, due to the lack of a reference corpus. Freitas and Quental [42] evaluates manually a random sample of the extracted relations, following a rating system for the relations (3: correct; 2: partially correct; 1: correct in general terms; 0: wrong).

An obstacle for adequate evaluation of Relation Extraction works in Portuguese lies in the comparison of the results, given that most RE researches for Portuguese use different resources, such as corpus and parser. In Table 5, we present the corpora used, the method applied, the respective evaluation method and corresponding performance of those works (refer to Sect. 3.3 for more details).

Although the values are not strictly comparable, we can see from Tables 4 and 5 that English systems, in general, report higher score values for this kind of task. We can also see that there is a large number of rule based systems. Besides those that use HAREM data, each has a different data collection.

6 Concluding remarks

The main contribution of this paper is to present a review of the current research in RE, addressing the progress and difficulties of the area. To our knowledge, no previous work presented such a complete overview considering the case of Portuguese language, a less resourceful language, in this scenario.

As seen throughout this paper, there are plenty of research efforts put into the development of better systems for the RE task. Many different computational approaches were tried, and many linguistic resources have been considered as an aid for the solution of this problem. Also, there is a considerable variety in the way these systems are evaluated.

For the classical and traditional supervised approaches, we see how hard it is to extend them for supporting new types of relations and entities, due to the need of specialized annotated data. While this area is heavily dependent of proper tagging, annotated corpora with such semantic information are particularly scarce and very costly. Even if these data are getting more fully available for standard entities such as *Person*, *Organization* and *Location*, they are still rare when the goal is extending the NER to new types of entities. However, such supervised approaches are still useful for specific domains—for example, biomedicine [66], where entities types are more regular.

But in the general picture, what we see is an increasing number of relation types being tackled in the literature, each associated to different entity types. Due to the time-consuming and mistake-prone nature of manual annotation, semi-supervised techniques such as bootstrapping appeared as an alternative. Indeed, semi-super-vised approaches are more adequate for open domain Relation Extraction systems, since they are quite exible in regards of the quantity of input data and they can be more easily extended to deal with new relation types. Open IE approaches are essential when the number of relations of interest is massive or unknown. On the other hand, while these new techniques to deal with the problem are getting more sophisticated, and the variety of data considered increases, many of the evaluations in this line of work are isolated and seldom based on a rather small sample, as illustrated in Table 4. This is due to the unavailability of Gold Standards for the more ambitious systems—for instance, those which considers the web as corpus.

The result is that innovative research is being conducted outside the frame of the quite long tradition of evaluation contests in this area, with the disadvantage of weaker evaluation methodologies. On the other hand, while this tradition in evaluation is seen for the English language (the first evaluation dates from 1995), this is not the case for Portuguese. This language was not included in any of the aforementioned evaluations, while languages such as Chinese and Spanish have been considered more recently. Specially for the Portuguese language, two evaluation contests were undertaken for similar problems, First HAREM and Second HAREM. The result is that the availability of resources are much less widespread for this language. Thus, shared evaluation of Relation Extraction systems for Portuguese is still in its infancy.

A joint evaluation effort for RE in Portuguese occurred just once, in the ReRelEM track of the Second HAREM, in which only three systems participated. Given the diversity on NEs and relations explored in this contest, they could not be fairly compared. Portuguese systems are still struggling to cope with the state-of-the-art methodologies, even though important resources were made available in these efforts, which greatly help in the research progress for this language [44].

Another general problem of the area is the fact that most methods need some sort of pre-processed data such as POS tags, parse trees, dependency parse trees and others. Thus, the pre-processing step is also prone to mistakes and might affect

the systems performance. There are fewer of those resources available for less technologically developed languages, such as Portuguese, usually with lower performance then the state-of-the-art performance of systems for languages as English. Thus, this situation brings even more challenges. Note that we might also face some problems regarding the changes that occur in the real world in RE task, such as employees changing employer or a company launching new products [31].

Besides, since the languages are different and the approaches towards each of them also differ greatly—from the data employed to the method itself—it is hard to compare RE systems for Portuguese with those for English. One way to reduce the impact of the lack of natural language resources for certain languages may be exploring the parallel corpora to transfer information from one language to another. Information transfer between languages can be very useful when the "donor" language (most of the time English) has more resources than the receiving one [106].

The study of this domain is of great importance for several NLP and business applications. Following the initial analysis made on [71], which focuses on kernel methods only, our review is the first to give a comprehensive analysis of the area, considering also resources and evaluation differences among the referred works. Our review also situates the Portuguese language in the picture of NLP research. We believe that it will serve as a helpful aid for researchers in the area.

Acknowledgments We thank the following people for their help in reviewing this article: Carolina de Abreu Pereira, Aline A. Vanin, and Cássia Trojahn. We thank the Brazilian funding agency FAPERGS/CAPES for the scholarship granted.

References

1. Agichtein E, Gravano L, SNOWBALL (2000) Extracting relations from large plain-text collections. In: 5th ACM international conference on digital libraries, pp 85–94
2. Bach N, Badaskar S (2007) A survey on relation extraction. Technical report, Literature review for language and statistics II 2007, Carnegie Mellon University
3. Banko M, Cafarella MJ, Soderl S, Broadhead M, Etzioni O (2007) Open information extraction from the web. In: IJCAI, pp 2670–2676 (2007)
4. Banko M, Etzioni O (2008) The tradeoffs between open and traditional relation extraction. In: McKeown K, Moore JD, Teufel S, Allan J, Furui S (eds) ACL. The Association for Computer, Linguistics, Bulgaria, pp 28–36
5. Bellare K, Mccallum A (2007) Learning extractors from unlabeled text using relevant databases. In: Sixth international workshop on information integration on the web (II Web)
6. Bick E (2000) The parsing system Palavras. In: Automatic grammatical analysis of Portuguese in a constraint grammar framework. University of Arhus, Arhus
7. Bick E (2003) Multi-level NER for Portuguese in a cg framework: PROPOR 2003. In: Lecture notes in computer science, vol 2721. Springer, Faro, pp 118–125
8. Bick E, Santos D, Afonso S, Marchi R (2007) Floresta sintactica: fico ou realidade. In: Santos D (ed) Avaliação Conjunta: Um novo paradigma no processamento computacional da lngua portuguesa, Chap 24. IST Press, Hamilton, pp 291–300 (2007)
9. Bollacker KD, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD conference, pp 1247–1250
10. Brin S (1998) Extracting patterns and relations from the World Wide Web. In: Atzeni P, Mendelzon AO, Mecca G (eds) WebDB. Lecture notes in computer science, vol 1590. Springer, Berlin, pp 172–183
11. Brucksen M, Souza JGC, Vieira R, Rigo S (2008) Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. In: Mota C, Santos D (eds) Segundo HAREM, Chap 14. Linguateca, Portuguese, pp 247–260
12. Buchholz S, Green D (2006) Quality control of treebanks: documenting, converting, patching. In: Calzolari N, Choukri K, Gangemi A, Maegaard B, Mariani J, Odjik J, Tapias D (eds) LREC 2006, pp 26–31
13. Bunescu RC, Mooney RJ (2007) Learning to extract relations from the Web using minimal supervision. ACL, Bulgaria
14. Cardoso N (2006) Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Master's thesis, Faculdade de Engenharia da Universidade do Porto, Porto
15. Cardoso N (2008) Rembrandt—reconhecimento de entidades mencionadas baseado em relações análise detalhada do texto. In: Mota C, Santos D (eds) Segundo HAREM, Chap 11. Linguateca, Portuguese, pp 195–211
16. Cardoso, N (2012) Rembrandt—a named-entity recognition framework. In: Chair NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Odijk J, Piperidis S (eds) Proceedings of the eight international conference on language resources and evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul
17. Carlson A, Betteridge J, Hruschka Jr ER, Mitchell TM (2009) Coupling semi-supervised learning of categories and relations. In: SemiSupLearn '09: proceedings of the NAACL HLT 2009 workshop on semi-supervised learning for natural language processing. Association for Computational Linguistics, Morristown, pp 1–9
18. Carvalho P, Oliveira HG, Mota C, Santos D, Freitas C (2008) Segundo HAREM: nodelo geral, novidades eavaliação. In: Mota C, Santos D (eds) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM
19. Chaves MS (2008) Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo HAREM. In: Mota C, Santos D (eds) Segundo HAREM, Chap 13. Linguateca, Portuguese, pp 231–245
20. Chaves MS, Silva MJ, Martins B (2005) A geographic knowledge base for semantic web applications. In: Heuser CA (ed) 20th Brazilian symposium on databases, pp 40–54
21. Chen Y, Zheng Q, Wang W, Chen Y (2010) Knowledge element relation extraction using conditional random fields. In: CSCWD, pp 245–250
22. Christensen J, Mausam, Soderland S, Etzioni O (2010) Semantic role labeling for open information extraction. In: Proceedings of the NAACL HLT 2010 first international workshop on formalisms and methodology for learning by reading. FAM-LbR '10. Association for Computational Linguistics, Stroudsburg, pp 52–60
23. Christensen J, Mausam, Soderland S, Etzioni O (2011) An analysis of open information extraction based on semantic role labeling. In: K-CAP, pp 113–120
24. Collovini S, Grando F, Souza M, Freitas L, Vieira R (2011) Semantic relations extraction in the organization domain. In: Proceedings of IADIS international conference on applied computing, Rio de Janeiro, pp 99–106

25. Culotta A, McCallum A, Betz J (2006) Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the main conference on human language technology conference of the North American chapter of the Association of Computational Linguistics. HLT-NAACL '06, Association for Computational Linguistics, Stroudsburg, pp 296–303 (2006)
26. Culotta A, Sorensen J (2004) Dependency tree kernels for relation extraction. In: Proceedings of the 42nd meeting of the association for computational linguistics (ACL'04), main volume, Barcelona, pp 423–429
27. Cunningham H, Maynard D, Bontcheva K, Tablan V (2002) GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia (2002)
28. Curran JR, Murphy T, Scholz B (2007) Minimising semantic drift with mutual exclusion bootstrapping. In: Proceedings of the conference of the Pacific Association for Computational Linguistics, pp 172–180
29. Day D, Aberdeen J, Hirschman L, Kozierok R, Robinson P, Vilain M (1997) Mixed-initiative development of language processing systems. In: Proceedings of the fifth conference on applied natural language processing
30. Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R (2004) The automatic content extraction (ACE) program: tasks, data, and evaluation. In: Lino MT, Xavier MF, Ferreira F, Costa R, Silva R (eds) Proceedings of the 4th international conference on language resources and evaluation—LREC 2004, Lisboa, pp 837–840
31. Drury, B.: A text mining system for evaluating the stock market's response to news. Ph.D. thesis, Faculdade de Cincias da Universidade do Porto (FCUP)
32. Ebecken NFF, Lopes MCS, de Arago Costa MC (2005) Minerao de textos. In: Rezende SO (ed) Sistemas inteligentes: fundamentos e aplicaes, Chap 13. Manole, Barueri, pp 337–370
33. Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A (2005) Unsupervised named-entity extraction from the web: an experimental study. Artif Intell 165(1):91–134
34. Etzioni O, Cafarella MJ, Downey D, Kok S, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A (2004) Web-scale information extraction in knowitall: preliminary results. In: WWW, pp 100–110
35. Etzioni O, Fader A, Christensen J, Soderland S, Mausam (2011) Open information extraction: the second generation. In: Twenty-second international joint conference on artificial intelligence, IJCAI, pp 3–10
36. Fader A, Soderland S, Etzioni O (2011) Identifying relations for open information extraction. In: EMNLP, pp 1535–1545
37. Fellbaum C (ed) WordNet an electronic lexical database. The MIT Press, Cambridge (1998)
38. Fernandes WPD, Motta E, Milidi L (2011) Quotation extraction for Portuguese. In: 8th Brazilian symposium in information and human language technology—STIL'2011, Cuiab, pp 204–208
39. Ferreira L, Oliveira C, Teixeira A, Cunha J (2009) Extração de informação de relatrios mdicos. Linguamatica 1(1):89–101
40. Freitag D, Mccallum A (2000) Information extraction with HMM structures learned by stochastic optimization. In: Proceedings of the seventeenth national conference on artificial intelligence. AAAI Press, Menlo Park, pp 584–589
41. Freitas C, Mota C, Santos D, Oliveira HG, Carvalho P (2010) Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the seventh international conference on language resources and evaluation (LREC'10). European Language Resources Association (ELRA), Valletta
42. Freitas C, Quental V (2007) Subsídios para a elaboração automática de taxonomias. In: V Workshop em Tecnologia da Informação e da Linguagem Humana—TIL 2007. Rio de Janeiro, Brazil
43. Freitas C, Rocha P, Bick E (2008) Floresta sintá(c)tica: bigger, thicker and easier. In: Proceedings of the 8th international conference on computational processing of the Portuguese language. PROPOR '08. Springer, Berlin, pp 216–219
44. Freitas C, Santos D, Mota C, Oliveira HG, Carvalho P (2009) Detection of relations between named entities: report of a shared task. In: EW-2009—semantic evaluations: recent achievements and future directions (NAACL-HLT 2009 workshop), Colorado
45. Freitas C, Santos D, Oliveira HG, Carvalho P, Mota C (2008) Relações semnticas do ReRelEM: alémdas entidades no Segundo HAREM, Chap 4. Linguateca, pp 75–94
46. Freitas C, Santos D, Oliveira HG, Quental V (2010) Varra: validação, avaliaçõo e revisão de relações semânticas no AC/DC, ELC
47. Gabbard R, Freedman M, Weischedel RM (2011) Coreference for learning to extract relations: yes virginia, conference matters. In: Proceedings of the 49th annual meeting of the ACL: short papers. The Association for Computer Linguistics, Portland, pp 288–293
48. 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, pp 10–18
49. Gonçalo Oliveira H, Gomes P, Onto PT (2011) Construção automática de uma ontologia lexical para o português. In: Luís, AR (ed) Estudos de Linguística, vol 1. Imprensa da Universidade de Coimbra, Coimbra
50. Gonçalo Oliveira H, Gomes P (2012) Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In: Natural language processing and information systems, 17h NLDB, vol 7337. LNCS. Springer, Groningen
51. Gonalo Oliveira H, Prez LA, Gomes P (2012) Exploring Onto.pt. In: Demo session of PROPOR 2012, 10th international conference on the computational processing of the Portuguese, language
52. Gruber TR (1992) Ontolingua: a mechanism to support portable ontologies. Technical report
53. GuoDong Z, Jian S, Jie Z, Min Z (2005) Exploring various knowledge in relation extraction. In: Proceedings of the 43rd annual meeting on Association for Computational Linguistics. ACL '05, Association for Computational Linguistics, Stroudsburg, pp 427–434
54. Hasegawa T, Sekine S, Grishman R (2004) Discovering relations among named entities from large corpora. In: ACL '04: proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, p 415
55. Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on computational linguistics, vol 2. COLING '92, Association for Computational Linguistics, Stroudsburg, pp 539–545
56. Hearst MA (1998) Automated discovery of wordNet relations. In: Fellbaum C (ed) WordNet: an electronic lexical database and some of its applications. MIT Press, Massachusetts, pp 1–26
57. Hobbs JR, Appelt D, Bear J, Israel D, Kameyama M, Stickel M, Tyson M (1997) Fastus: a cascaded finite-state transducer for extracting information from natural-language text. In: Roche E, Schabes Y (eds) Finite-state language processing. MIT Press, Cambridge, pp 383–406
58. Hoffmann R, Zhang C, Ling X, Zettlemoyer LS, Weld DS (2011) Knowledge-based weak supervision for information extraction of overlapping relations. ACL, Stroudsburg, pp 541–550

59. Ji H, Grishman R (2006) Analysis and repair of name tagger errors. In: ACL 2006, 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics. The Association for Computer Linguistics, Sydney
60. Jiang J (2012) Information extraction from text. In: Mining text data, pp 11–41
61. Kambhatla N (2004) Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004 on interactive poster and demonstration sessions. ACL demo '04, Association for Computational Linguistics, Stroudsburg
62. Krestel R, Bergler S, Witte R (2008) Minding the source: automatic tagging of reported speech in newspaper articles. LREC
63. Lafferty JD, McCallum A, Pereira FCN (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 Publishers Inc., San Francisco, pp 282–289
64. Li H, Bollegala D, Matsuo Y, Ishizuka M (2011) Using graph based method to improve bootstrapping relation extraction. In: CICLing, vol 2, pp 127–138
65. Li Y, Jiang J, Chieu HL, Chai KMA (2011) Extracting relation descriptors with conditional random fields. In: Proceedings of 5th international joint conference on natural language processing. Asian Federation of Natural Language Processing, Chiang Mai, pp 392–400
66. Liu Y, Shi Z, Sarkar A (2007) Exploiting rich syntactic information for relation extraction from biomedical articles. In: Human language technologies 2007: the conference of the North American chapter of the Association for Computational Linguistics; companion volume, short papers. NAACL-Short '07, Association for Computational Linguistics, Stroudsburg, pp 97–100
67. Marcus MP, Santorini B, Marcinkiewicz MA (1993) Building a large annotated corpus of english: the Penn Treebank. Comput Linguist 19(2):313–330
68. Marrafa P (2000) Portuguese wordnet: general architecture and internal semantic relations. DELTA 18:131–146
69. McIntosh T, Curran JR (2009) Reducing semantic drift with bagging and distributional similarity. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. Association for Computational Linguistics, Suntec, pp 396–404
70. Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant supervision for relation extraction without labeled data. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, ACL '09, vol 2. Association for Computational Linguistics, Stroudsburg, pp 1003–1011
71. Moncecchi G, Minel JL, Wonsever D (2010) A survey of kernel methods for relation extraction. In: Workshop on natural language processing and web-based technologies in conjunction with IBERAMIA
72. MUC-6 (1995) Conference task definition. In: Proceedings of the sixth message understanding conference—MUC-6
73. MUC-7 (1997) Conference task definition. In: Proceedings of the seventh message understanding conference—MUC-7
74. Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Linguist Investig 30(1):3–26
75. NIST, ACE (2008) Automatic content extraction 2008 evaluation plan (ace08). Technical report, NIST
76. Oliveira HG, Santos D, Gomes P, Seco N (2008) Papel: a dictionary-based lexical ontology for Portuguese. In: Teixeira A, de Lima VLS, de Oliveira LC, Quaresma P (eds) Computational processing of the Portuguese language, 8th international conference, PROPOR 2008. Lecture notes in computer science, vol 5190. Springer, Berlin, pp 31–40
77. OpenNLP (2010) open-source framework to develop natural language applications. http://opennlp.apache.org/. Accessed 16 June 2012
78. Palmer DD, Day DS (1997) A statistical profile of the named entity task. In: ANLP, pp 190–193
79. Pantel P, Pennacchiotti M (2006) Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: International conference on computational linguistics/Association. ACL Press, Sydney, pp 113–120
80. Pouliquen B, Steinberger R, Best C (2007) Automatic detection of quotations in multilingual news. In: International conference recent advances in natural language processing (RANLP 2007). Borovets, Bulgaria, pp 487–492
81. Qian L, Zhou G, Kong F, Zhu Q, Qian P (2008) Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In: COLING, pp 697–704
82. Ruiz-Casado M, Alfonseca E, Castells P (2007) Automatising the learning of lexical patterns: an application to the enrichment of wordnet by extracting semantic relationships from wikipedia. Data Knowl Eng 61(3):484–499
83. Sahay S, Lee J, Krishnamurthi N (2008) Relationship extraction from biomedical documents using conditional random fields. Technical report, College of Computing, Georgia Institute of Technology
84. dos Santos CN, Oliveira C (2005) Aplicação de aprendizado baseado em transformação na identificação de sintagmas nominais. In: Anais do XXV Congresso da SBC, workshop on technology on information and human language (TIL'05), Brazil, pp 2138–2147
85. Santos D, Mamede N, Baptista J (2010) Extraction of family relations between entities. In: Barbosa LS, Correia MP (ed) Proceedings of the INForum 2010—II Simpsio de Informtica. Braga, Portugal, pp 549–560
86. Santos D (2007) Avaliação conjunta. In: Santos D (ed) Avaliação Conjunta: Um novo paradigma no processamento computacional da língua portuguesa, chap 1. IST Press, Lisboa, pp 1–12
87. Santos D, Barreiro A, Freitas C, Oliveira HG, Medeiros JC, Costa L, Gomes P, Silva R (2010) Relações semânticas em português: comparando o tep, o mwn.pt, o port4nooj e o papel. In: Brito AM, Silva F, Veloso J, Fiéis A (eds) Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística, APL, Portuguese, pp 681–700
88. Santos D, Cardoso N (2007) Breve introdução ao HAREM. In: Santos D, Cardoso N (eds) Econhecimento de entidades mencionadas em português: documentação e actas do HAREM, a primeira avaliação conjunta na área, Chap 1. Linguateca, Portuguese, pp 1–16
89. Sarawagi S (2008) Information extraction. Found Trends Databases 1(3):261–377
90. Sarmento L, Nunes S (2009) Automatic extraction of quotes and topics from news feeds. In: 14th doctal symposium on informatics, engineering
91. Dias-da Silva BC, Oliveira MF, Moraes HR, Hasegawa R, Amorim D, Paschoalino C, Nascimento AC (2000) Construo de um thesaurus eletrnico para o portugus do brasil. V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR 2000, vol 4, pp 1–11
92. Dias-da Silva BC, Felippo AD, Hasegawa R (2006) Methods and tools for encoding the wordnet.br sentences, concept glosses, and conceptual-semantic relations. In: Vieira R, Quaresma P, das Graas Volpe Nunes M, Mamede NJ, Oliveira C, Dias MC (eds) PROPOR. Lecture notes in computer science, vol 3960. Springer, Berlin, pp 120–130

93. Specia L, Motta E (2006) A hybrid approach for extracting semantic relations from texts. In: Proceedings of the 2nd workshop on ontology learning and population: bridging the gap between text and knowledge. Association for Computational Linguistics, Sydney, pp 57–64
94. Sun A (2009) A two-stage bootstrapping algorithm for relation extraction. In: Proceedings of RANLP 2009—recent advances in natural language processing. Borovets, Bulgaria
95. Sun A, Grishman R, Sekine S (2011) Semi-supervised relation extraction with large-scale word clustering. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. HLT '11, pp 521–529. Association for Computational Linguistics, Stroudsburg
96. Taba LS, de Medeiros Caseli H (2012) Automatic hyponymy identification from Brazilian Portuguese texts. In: de Medeiros Caseli H, Villavicencio A, Teixeira AJS, Perdigo F (eds) PROPOR. Lecture notes in computer science, vol 7243. Springer, Berlin, pp 186–192
97. Tanev H, Zavarella V, Linge J, Kabadjov M, Piskorski J, Atkinson M, Steinberger R (2009) Exploiting machine learning techniques to build an event extraction system for Portuguese and Spanish. Linguamatica 1(2):55–66
98. Wang T, Li Y, Bontcheva K, Cunningham H, Wang J (2006) Automatic extraction of hierarchical relations from text. In: Proceedings of the 3rd European conference on the semantic web: research and applications. ESWC'06. Springer, Berlin, pp 215–229
99. Wu F (2010) Machine reading: from wikipedia to the web. Master's thesis. University of Washington, Seattle
100. Wu F, Weld DS (2010) Open information extraction using wikipedia. In: ACL, Stroudsburg, pp 118–127
101. Xavier CC, de Lima VLS (2010) A semi-automatic method for domain ontology extraction from Portuguese language wikipedia's categories. In: SBIA, pp 11–20
102. Yates A, Banko M, Broadhead M, Cafarella MJ, Etzioni O, Soderland S (2007) Textrunner: open information extraction on the web. In: HLT-NAACL (demonstrations), pp 25–26
103. Zelenko D, Aone C, Richardella A (2003) Kernel methods for relation extraction. J Machine Learn Res 3:1083–1106
104. Zhao S, Grishman R (2005) Extracting relations with integrated information using kernel methods. ACL. The Association for Computer Linguistics, Stroudsburg
105. Zhu J, Nie Z, Liu X, Zhang B, Wen JR (2009) Statsnowball: a statistical approach to extracting entity relationships. In: 18th international conference on world wide web. ACM, New York, pp 101–110
106. Zitouni I, Florian R (2008) Mention detection crossing the language barrier. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP)