Semantic Relation Extraction. Resources, Tools and Strategies*

Marcos Garcia

Grupo LyS, Departamento de Galego-Português, Francês e Linguística
Universidade da Coruña, Campus da Coruña
marcos.garcia.gonzalez@udc.gal

NOTICE: this is the author version of a paper that was accepted for publication in *Computational Processing of the Portuguese Language*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. A definitive version has been published in *Computational Processing of the Portuguese Language*, Volume 9727 of the series Lecture Notes in Computer Science (pp 141–152), available at: http://link.springer.com/chapter/10.1007/978-3-319-41552-9_15

Abstract. Relation extraction is a subtask of information extraction that aims at obtaining instances of semantic relations present in texts. This information can be arranged in machine-readable formats, useful for several applications that need structured semantic knowledge. The work presented in this paper explores different strategies to automate the extraction of semantic relations from texts in Portuguese, Galician and Spanish. Both machine learning (distant-supervised and supervised) and rule-based techniques are investigated, and the impact of the different levels of linguistic knowledge is analyzed for the various approaches. Regarding domains, the experiments are focused on the extraction of encyclopedic knowledge, by means of the development of biographical relations classifiers (in a closed domain) and the evaluation of an open information extraction tool. To implement the extraction systems, several natural language processing tools have been built for the three research languages: From sentence splitting and tokenization modules to part-of-speech taggers, named entity recognizers and coreference resolution systems. Furthermore, several lexica and corpora have been compiled and enriched with different levels of linguistic annotation, which are useful for both training and testing probabilistic and symbolic models. As a result of the performed work, new resources and tools are available for automated processing of texts in Portuguese, Galician and Spanish.

Keywords: information extraction, natural language processing, named entity recognition, part-of-speech tagging, coreference resolution

* This work has been partially supported by the Spanish Ministry of Economy and Competitiveness through the project FFI2014-51978-C2-1-R, and by a *Juan de la Cierva formación* grant, reference FJCI-2014-22853.

1 Introduction

In recent years, the amount of data generated by our society increased exponentially, and several studies show that this growth is currently even faster. An important portion of these data is compound by text, published every day in digital media and different languages.

This huge amount of text contains information that can be very useful for various applications in many areas. However, the size of these data makes impossible their processing through reading.

Aimed at simplifying this process, our work has the main objective of developing linguistically-based resources and tools for the automatic extraction of semantic information as well as their evaluation. The extraction of structured semantic knowledge from free text is useful both for theoretical purposes (it gives information about how semantic relations are represented linguistically) and from a pragmatic point of view (it permits to create structured databases and other useful resources) [19].

Thus, the work carried out involved the development and evaluation of Relation Extraction (RE) tools, capable of automatically extracting semantic knowledge from free text in Portuguese, Galician and Spanish. As an example, from a sentence (in Portuguese) like the following:

> John A. Garcia (nascido em 1949 na Galiza) é um dos pioneiros da indústria moderna americana de videojogos e o atual presidente da Novalogic.[1]

a semantic RE system could extract the following structured knowledge:

1. John A. Garcia *BirthDate* 1949
2. John A. Garcia *BirthPlace* Galiza
3. John A. Garcia *PresidentOf* Novalogic

Taking the above into account, we explore various strategies for building RE systems: From symbolic methods relying on a syntactic analysis to different machine learning models that use both weakly-supervised and supervised approaches. Furthermore, in order to implement the RE systems, it was necessary to build and adapt several tools for performing automatic linguistic analysis in the target languages: From modules for sentence boundary identification, morphological analyzers and lemmatizers (that recognize, for instance, *nascido* as a form of the verb *nascer*), to named entity recognizers (that can identify *John A. Garcia* as a single personal name, or *Galiza* as a location) or syntactic parsers.

The results of the performed work bring interesting information about the use of linguistic data in different strategies for RE in Portuguese, Galician, and Spanish. On the one hand, several evaluations showed that linguistic information

[1] A possible English translation could be: "John A. Garcia (born in 1949 in Galicia) is one of the pioneers of the modern American computer game industry and the current president of Novalogic.".

(namely those produced by lemmatization and semantic classification) is critical for building machine learning systems for relation extraction, even though their performance could be negatively affected if the linguistic analysis produces errors.

On the other hand, rule-based approaches obtained high-quality results in some tasks. For instance, in Named Entity Recognition (NER) —with competitive results when compared to a supervised approach—, or in relation extraction, with precision results between 85% and 95% (depending on the relation and language).

Also, we introduce novel strategies for improving tasks such as PoS-tagging (using dedicated parsers), corpus annotation (by means of distant-supervision) and Open Information Extraction (OIE), using automatic Coreference Resolution (CR). Finally, the performed work makes freely available a large set of resources and tools for the three target languages.[2]

Apart from this introduction, this paper is organized as follows: Section 2 presents the main hypotheses and objectives of our work as well as the methodological aspects. Then, the structure is outlined in Section 3, which also shows some of the main experiments and results. After that, Section 4 includes an overview of some of the related work for each of the performed tasks, while the conclusions are shown in Section 5.

2 Problems, Hypotheses, and Objectives

Apart from the main problem presented in the introduction (i.e., the difficulty of taking advantage of the vast amount of data that is being produced), two other related issues need to be solved in order to face the mentioned one:

- Lack of resources and tools for RE in Portuguese, Galician and Spanish.
- Need for multilingual Natural Language Processing (NLP) tools and resources for text processing (previous to the extraction).

Aimed at facing these problems, three main hypotheses were formulated, presented here as questions:

- Is it possible to create RE systems capable of extracting accurate biographical knowledge in the three target languages?
- What kind of linguistic information is needed to build symbolic and statistical NLP tools?
- Is it feasible to develop rapidly and to adapt NLP resources and tools that are needed for performing the extractions?

During the work carried out, some other issues were taken into account apart from the mentioned ones. Problems such as the PoS-tagging of several varieties (national and orthographic) of Portuguese, the correction of the critical

[2] All of them are freely available at `http://gramatica.usc.es/~marcos/phd.html`

PoS-tagging errors or the best combination of coreference resolution with open information extraction, among others.

Once formulated the different hypotheses, a primary objective was defined as follows:

– The main goal consists in evaluating different strategies for the extraction of encyclopedic knowledge —mainly biographic— in closed domain, in Portuguese, Galician and Spanish.

To achieve the primary objective, several parallel goals were defined, that can be summarized in a single one:

– The implementation or adaptation of the NLP tools that are required for semantic RE in Portuguese, Galician, and Spanish.

These objectives are gradually achieved during our work, aimed at answering the formulated hypotheses and solving the problems that had been found. In this respect, we introduce the tools and resources that have been developed and their evaluation, as well as the various strategies for semantic RE that have been implemented.

2.1 Methodology

From a methodological point of view, our work is mainly based on the use of linguistic knowledge for NLP, but it combines this information with approaches from other areas (such as machine learning) in order to better achieve the proposed objectives. Thus, during the implementation and adaptation of each tool and resource, the theoretical proposals were taken into account, but the quality of the results was prioritized over the formal consistency. Therefore, the work is essentially pragmatic.

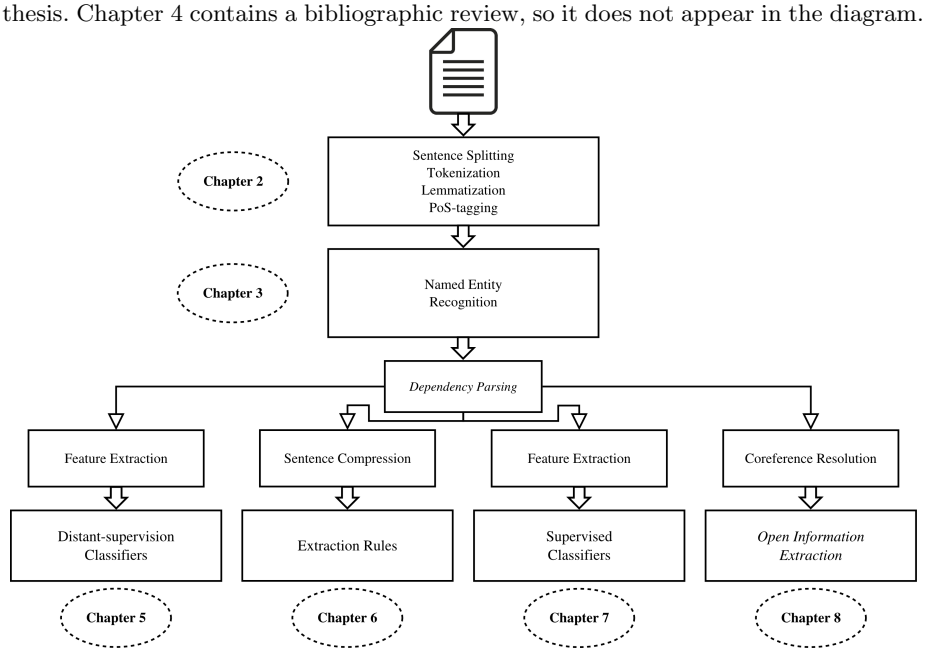
3 Structure, experiments and results

Our work can be divided into three well-distinguished parts: (i) natural language processing before the extraction, (ii) strategies for relation extraction and (iii) the combination of coreference resolution with open information extraction. Regarding the chapter structure, it is summarized in Figure 1.

3.1 NLP before the extraction

Before the implementation of RE systems, we adapted and the evaluated different modules for tokenization, sentence identification, lemmatization and morphosyntactic analysis for Portuguese and Galician [20], as well as a strategy for PoS-tagging correction [21]. We also made and adapted several tagged corpora and morphosyntactic dictionaries for different linguistic varieties.

Fig. 1. Diagram of the processes carried out in each chapter. The input (top) is plain text, and there are evaluated four different strategies of information extraction (bottom) in chapters 5 to 8. Elements in italic were not implemented specifically in the thesis. Chapter 4 contains a bibliographic review, so it does not appear in the diagram.



Specially for Portuguese, we addressed the problem of PoS-tagging different varieties of this language. Several lexica and corpora has been combined to train morphosyntactic analyzers for tagging texts in various national varieties (Brazil, Portugal, Angola, etc.) and spelling systems (before and after the *Acordo Ortográfico de 1990*) [30].

Different experiments showed that the resulting tools achieve competitive performance on different tasks, with sentence splitting, tokenization and lemmatization results between $> 98\%$ and $> 99\%$, depending on the language.

Concerning PoS-tagging, the evaluation in different varieties of Portuguese showed that a single model (e.g., European Portuguese) trained with consistent resources can achieve better performance than complex models that combine corpora from different varieties.

Furthermore, we also presented a set of tools (both statistical and based on linguistic rules and resources) for named entity recognition, capable of identifying and semantically classifying named entities in free text automatically [16, 31]. The presented tools, which work in the three referred languages, recognize proper nouns such as persons, organizations, and locations, as well as other expressions such as dates, quantities, etc.

Named entity recognition was approached in two steps: the identification of the named entity boundaries (with F1 results of 91% in Portuguese and 94%

in Galician) and the semantic classification of each entity (≈ 75% F1 in both Portuguese and Galician). Several experiments carried out showed that the supervised classifier achieves better results when evaluated in corpora from the same domain as the training data. However, the resource-based method, which takes advantage of large lists of gazetteers, obtains more stable values across different domains.

3.2 Relation extraction strategies

Three different strategies were evaluated for performing relation extraction in a closed domain: (i) a distant-supervision approach, (ii) the use of supervised classifiers and (iii) a novel rule-based strategy that takes advantage of text compression techniques.

First we implemented and evaluated strategies for distant-supervision RE [23, 25]. This technique allows the user to obtain labeled data in a semi-automatic way, using semantically related pairs extracted from structured resources. These data are then used for building machine learning models that are capable of automatically extracting new knowledge.

The evaluation was performed with the relation `Profession` in Portuguese and Spanish, obtaining competitive results (78% – 83% F1 in Portuguese and Spanish, respectively) when generalizing the patterns by means of the *longest common string algorithm*.

Then, we evaluated the impact of different linguistic knowledge in the training of supervised classifiers for RE [26]. It was performed a detailed analysis of several sets of classifiers that only differ in the linguistic knowledge they use: From basic lexical units to deep syntactic information, combined with semantic and pseudo-syntactic knowledge. Two corpora (one for Portuguese and other for Spanish) were semi-automatically labeled with five biographical relations using a distant-supervision approach. After that, the annotation was manually corrected, and different evaluations showed the benefits of lemmatization and NER for biographical RE.

Finally, we developed a new technique that consists in the application of text compression methods for simplifying the contexts containing semantic relations [22, 24]. This method is combined with a strategy, inspired in distant-supervision, for semi-automatic building of lexico-syntactic rules for RE. The performed experiments showed that the proposed method keeps the high-precision values of the pattern-matching approaches while increasing the recall. Using both encyclopedic and journalistic corpora, the systems extracted tens of thousands of semantically related pairs, with precision values between 85% and 96%, depending on the language, the domain, and the relation.

3.3 Coreference resolution and open information extraction

After evaluating closed domain RE, we studied the impact of CR in open information extraction in the three target languages. With the mentioned objective, it was developed a deterministic system for automatic CR of person entities in

Table 1. Examples of open information extraction without coreference resolution (OIE) and in combination with CR (CR+OIE).

Sentence (Spanish)	"Debutó en la Tercera división"
Sentence (English)	"[-] debuted in the third division"
OIE Extraction	∅
CR+OIE Extract. (Sp)	Ander Herrera *debutó_en* la Tercera división
CR+OIE Extract. (Eng)	Ander Herrera *debuted_in* the third division
Sentence (Portuguese)	"Anderson viajou por Europa"
Sentence (English)	"Anderson traveled in Europe"
OIE Extraction (Port)	Anderson *viajou_por* Europa
OIE Extraction (Eng)	Anderson *traveled_in* Europe
CR+OIE Extract. (Port)	Wes Anderson *viajou_por* Europa
CR+OIE Extract. (Eng)	Wes Anderson *traveled_in* Europe

Portuguese, Galician, and Spanish. The tool, evaluated in different corpora also created during our work [29], is capable of identifying and linking various expressions that refer to the same person in a text (*Miguel Gomes*, *the film director*, *he*, *Gomes*, etc.) [27].

Coreference resolution was evaluated together with *DepOE*, a multilingual tool for OIE also developed in our research group [17]. The performed tests show that the combination of these two methods allows the extraction to improve both its precision and its recall, being a promising strategy for further research [28]. In this regard, Table 1 exemplifies how a previous application of a CR tool improves the open information extraction output.

The results of this combination (Table 2) show the benefits of applying coreference resolution before open information extraction: on average, the number of extractions increased 22.7%, while the precision was 10.6% better. Finally, it was calculated an *enrichment* value as follows: we verified, from all the correct extractions, if the personal mention had been correctly solved by the CR tool. These cases were divided by the total number of correct extractions, being these results considered as the *enrichment* value. Although these results are not a direct evaluation of OIE, they suggest that the extraction is $\approx 79\%$ better when applied after CR.

Table 2. Results of two runs of the OIE system (OIE and CR+OIE) in the three target languages. *Prec.* means Precision, while the *Wikipedia* and *Journal* values are the number of extractions in these domains. *Enrich.* is the enrichment caused by the previous execution of the CR tool.

Language	**OIE Extraction**			**CR+OIE Extraction**			**Enrich.**
	Wikipedia	*Journal*	*Prec.*	*Wikipedia*	*Journal*	*Prec.*	
Portuguese	82	133	39%	111	155	56%	75%
Galician	168	114	49%	221	115	54%	77%
Spanish	47	82	49%	80	86	58%	84%

4 Related Work

This section briefly presents, following the structure presented in Section 3, some of the works that have been considered more relevant to each of the performed tasks.

Concerning the first steps of the NLP pipeline (sentence splitting, tokenization, lemmatization and PoS-tagging), some strategies already presented for Portuguese [5, 6] and Galician [32] were used in order to adapt the FreeLing suite [39] for these languages, following the EAGLES guidelines for morphosyntactic annotation [34].

After that, both probabilistic [10] and knowledge-based [16] models were used for implementing NER tools for Portuguese and Galician. NER was handled taking into account "timex", "numex" and "enamex" expressions as defined in the MUC-7 conference [36], thus differing from other approaches such as the HAREM evaluations [43, 38], which propose a more fine-grained entity classification.

About relation extraction, several works (such as [7]) were inspired by the pattern-matching approach used by Hearst [33], while some others proposed strategies for the automatic learning of new extraction patterns [1, 14, 41].

More recently, new strategies were used to both reduce the effort of annotating training data and increase the number of extracted relations. Thus, techniques such as distant-supervision take advantage of large repositories of structured data for automatically obtaining training examples [37].

As pointed out in the previous section, open information extraction [2] is a new paradigm that extracts triples (with the following structure: *argument_1* `verb-based_relation` *argument_2*) without the need of previously define the target relations. Some OIE approaches use training data and shallow parsing for building the extractor [15], while others rely on heuristics based on dependency parsing [12].

In Portuguese, three different systems were presented to the ReRelEM task [38], focused on generic relation extraction: REMBRANDT [9], which uses knowledge extracted from the Wikipedia, SEI-Geo [11], that applies patterns similar to those used by Hearst [33], and SeRELeP [8], a rule-based system which identifies relations relying on NER labels.

Finally, regarding coreference resolution, the tool implemented in our work was inspired by the entity-centric approach presented in [35], enriched with some linguistic knowledge such as the proposed in [40].

Also, we followed (with minor differences) the guidelines proposed in [42] for annotating three corpora with coreference information of person entities.

In general, the mentioned papers were carefully studied to build different resources and tools for the research languages, combining strategies proposed by different authors and adapting them to our objectives. Also, it is important to note that out work took advantage of several available resources [4, 13, 3] and tools [10, 18, 39] for different languages.

5 Conclusions

The work that we carried out permitted, on one hand, to develop and evaluate novel and promising techniques for different types of semantic relation extraction, both in a closed domain and using OIE. Several tests in various languages proved that it is feasible to obtain automatically structured knowledge from free text.

On the other hand, it has been shown the effectiveness of different strategies for the development and adaptation of resources and tools for NLP in the three target languages.

Taking into account that the presented work takes advantage of both theoretical and applied linguistics, some parts proved that the combination of linguistic knowledge with statistical methods is useful in the different NLP tasks that have been evaluated.

Several tests showed that linguistic information (especially those produced by lemmatization as well as by semantic classification) is essential for building machine learning classifiers for relation extraction.

Also, those approaches based on syntactic dependencies achieved high-quality results in some tasks, such as the combination of coreference resolution with OIE. The promising results of this strategy turn it an interesting approach to further work on open domain information extraction.

As it has been said, it was necessary to implement several NLP tools and resources for the three research languages. This way, some of the developed tools were the first ones for performing various NLP tasks in Galician, and some others were the first ones with open source licenses for Portuguese and Spanish.

5.1 Resources and Tools

Besides the theoretical conclusions, the contributions of our work also include the following resources and tools:

- Sentence splitting modules for Portuguese and Galician.
- Tokenization modules for Portuguese and Galician.
- Morphological analysis modules for Portuguese and Galician.
- Morphosyntactic analysis modules for Portuguese and Galician.
- Adaptation of Bosque 8.0 corpus tags to EAGLES standard.
- Adaptation of LABEL-LEX (SW) lexicon to EAGLES standard.
- PoS-tagged corpora for different varieties of Portuguese and Galician.
- Lexica (and extension of existing lexica) with PoS-tags for different varieties of Portuguese and Galician.
- Named entity identification modules for Portuguese and Galician.
- Named entity classification modules for Portuguese and Galician.
- Modules for recognizing numerical expressions, quantities and hours for Portuguese and Galician.
- Named entity annotation of Bosque 8.0 corpus.
- Testing corpus with named entity annotation for Galician.

- Corpora with annotation of biographical relations for Portuguese and Spanish.
- Coreference resolution tool for person entities for Portuguese, Galician and Spanish.
- Corpora with coreferential annotation of person entities for Portuguese, Galician and Spanish.

All the resources and tools are available under open source licenses (GPLv3) or keeping the original license in the case of adaptations.

References

1. Agichtein, E., Gravano, L.: Snowball: Extracting Relations from Large Plain-Text Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries. pp. 85–94 (2000)
2. Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI 2007). pp. 2670–2676 (2007)
3. Barcala, F.M., Domínguez Noya, E.M., Otero, P.G., López Martínez, M., Moscoso Mato, E.M., Rojo, G., Santalla del Río, M.P., Sotelo Docío, S.: A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a in a Minority Language. In: Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 3rd Language & Technology Conference. pp. 359–363 (2007)
4. Bosque 8.0: Uma floresta integralmente revista por linguistas (2008), `http://www.linguateca.pt/Floresta/corpus.html\#bosque`
5. Branco, A., Silva, J.: Contractions: breaking the tokenization-tagging circularity. In: Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence (LNCS/LNAI). vol. 2721, pp. 167–170. Springer-Verlag (2003)
6. Branco, A., Silva, J.: Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). pp. 507–510 (2004)
7. Brin, S.: Extracting patterns and relations from the World Wide Web. In: Proceedings of the WebDB Workshop at the 6th International Conference on Extending Database Technology (EDBT 1998). pp. 172–183 (1998)
8. Bruckschen, M., Camargo de Souza, J., Vieira, R., Rigo, S.: Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. In: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, chap. 14, pp. 247–260. Linguateca (2008)
9. Cardoso, N.: REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, pp. 195–211. Linguateca (2008)
10. Carreras, X., Márquez, L., Padró, L.: A simple named entity extractor using AdaBoost. In: Proceedings of the 7th Conference on Natural Language Learning at HLT/NAACL 2003. vol. 4, pp. 152–155. ACL (2003)

11. Chaves, M.: Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, pp. 231–245. Linguateca (2008)
12. Corro, L.D., Gemulla, R.: ClausIE: Clause-based Open Information Extraction. In: Proceedings of the 22nd international conference on World Wide Web (WWW 2013). pp. 355–366 (2013)
13. Eleutério, S., Ranchhod, E., Mota, C., Carvalho, P.: Dicionários Electrónicos do Português. Características e Aplicações. In: Actas del VIII Simposio Internacional de Comunicación Social. pp. 636–642 (2003)
14. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D., Yates, A.: Web-scale information extraction in KnowItAll. In: Proceedings of the 13th international conference on World Wide Web (WWW 2004). pp. 100–110. ACM (2004)
15. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam: Open information extraction: The second generation. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). pp. 3–10 (2011)
16. Gamallo, P., Garcia, M.: A resource-based method for named entity extraction and classification. In: Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence (LNCS/LNAI). vol. 7026/2011, pp. 610–623. Springer-Verlag (2011)
17. Gamallo, P., Garcia, M., Fernández-Lanza, S.: Dependency-based Open Information Extraction. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012). pp. 10–18. ACL (2012)
18. Gamallo, P., González López, I.: A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. International Journal of Corpus Linguistics 16(1), 45–71 (2011)
19. Garcia, M.: Extracção de Relações Semânticas. Recursos, Ferramentas e Estratégias. Ph.D. thesis, Universidade de Santiago de Compostela (2014)
20. Garcia, M., Gamallo, P.: Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. Linguamática. Revista para o Processamento Automático das Línguas Ibéricas 2(2), 59–67 (2010)
21. Garcia, M., Gamallo, P.: Using Morphosyntactic Post-Processing to Improve PoS-tagging Accuracy. In: Proceedings of the 9th International Conference on Computational Processing of Portuguese Language (PROPOR 2010). Ext. Act. (2010)
22. Garcia, M., Gamallo, P.: A Weakly-Supervised Rule-Based Approach for Relation Extraction. In: Proceedings of the XIV Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2011). Workshop on Knowledge Extraction and Exploitation from Semi-structures Online Sources (KEESOS) (2011)
23. Garcia, M., Gamallo, P.: An Exploration of the Linguistic Knowledge for Semantic Relation Extraction in Spanish. In: Proceedings of the Joint Workshop FAM-LbR/KRAQ'11. Learning by Reading and its Applications in Intelligent Question-Answering at 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). pp. 7–12 (2011)
24. Garcia, M., Gamallo, P.: Dependency-Based Text Compression for Semantic Relation Extraction. In: Proceedings of the Workshop on Information Extraction and Knowledge Acquisition (IEKA 2011) at 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011). pp. 21–28 (2011)
25. Garcia, M., Gamallo, P.: Evaluating Various Features on Semantic Relation Extraction. In: Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011). pp. 721–726 (2011)

26. Garcia, M., Gamallo, P.: Exploring the effectiveness of linguistic knowledge for biographical relation extraction. Natural Language Engineering 21(4), 519–551 (2013)
27. Garcia, M., Gamallo, P.: An entity-centric coreference resolution system for person entities with rich linguistic information. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 741–752 (2014)
28. Garcia, M., Gamallo, P.: Entity-centric coreference resolution of person entities for open information extraction. Procesamiento del Lenguaje Natural 53, 25–32 (2014)
29. Garcia, M., Gamallo, P.: Multilingual corpora with coreference annotation of person entities. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). pp. 3229–3233. ELRA (2014)
30. Garcia, M., Gamallo, P., Gayo, I., Pousada Cruz, M.: PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems. Procesamiento del Lenguaje Natural 53, 95–101 (2014)
31. Garcia, M., Gayo, I., González López, I.: Identificação e Classificação de Entidades Mencionadas em Galego. Estudos de Lingüística Galega 4, 13–25 (2012)
32. Graña, J., Barcala, F., Vilares, J.: Formal Methods of Tokenization for Part-of-Speech Tagging. In: Computational linguistics and intelligent text processing, vol. 2276/2002, pp. 123–144. Springer-Verlag (2002)
33. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics. vol. 2, pp. 539–545. ACL (1992)
34. Leach, G., Wilson, A.: Recommendations for the Morphosyntactic Annotation of Corpora. Tech. rep., Expert Advisory Group on Language Engineering Standard (EAGLES) (1996)
35. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4), 885–916 (2013)
36. Mikheev, A., Grover, C., Moens, M.: XML tools and architecture for Named Entity Recognition. Journal of Markup Languages: Theory and Practice 1(3), 89–113 (1998)
37. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009). pp. 1003–1011. ACL (2009)
38. Mota, C., Santos, D. (eds.): Desafios na avaliação conjunta do reconhecimento de entidades mencionadas. O Segundo HAREM. Linguateca (2008)
39. Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA (2012)
40. Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., Muñoz, R.: An algorithm for anaphora resolution in Spanish texts. Computational Linguistics 27(4), 545–567 (2001)
41. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: Proceedings of the International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics (COLING/ACL 2006). pp. 113–120. ACL (2006)
42. Recasens, M., Martí, M.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation 44(4), 315–345 (2010)
43. Santos, D., Cardoso, N. (eds.): Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca (2007)