Multilingual corpora with coreferential annotation of person entities

Marcos Garcia, Pablo Gamallo

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
marcos.garcia.gonzalez@usc.es, pablo.gamallo@usc.es

Abstract

This paper presents three corpora with coreferential annotation of person entities for Portuguese, Galician and Spanish. They contain coreference links between several types of pronouns (including elliptical, possessive, indefinite, demonstrative, relative and personal clitic and non-clitic pronouns) and nominal phrases (including proper nouns). Some statistics have been computed, showing distributional aspects of coreference both in journalistic and in encyclopedic texts. Furthermore, the paper shows the importance of coreference resolution for a task such as Information Extraction, by evaluating the output of an Open Information Extraction system on the annotated corpora. The corpora are freely distributed in two formats: (i) the SemEval-2010 and (ii) the *brat rapid annotation tool*, so they can be enlarged and improved collaboratively.

Keywords: coreference, anaphora, corpus annotation

1. Introduction

When a text is produced, several concepts are often expressed in many different ways without losing the reference to the same discourse entity. Thus, a person like "Ayrton Senna" may be referred by a personal pronoun ("He"), a noun phrase ("the Brazilian driver"), or a relative pronoun ("who") among other linguistic units. Different expressions referring to the same discourse entity are in a coreference relation (Recasens and Martí, 2010).

Knowing the behavior of this phenomenon allows us, from a linguistic point of view, to better understand how a discourse is organized at the semantico-referential level (Gordon and Hendrick, 1998). From the Natural Language Processing perspective, coreference resolution is a crucial task for different applications such as Text Summarization (Steinberger et al., 2007) or Information Extraction (Banko and Etzioni, 2008).

Information about people is one of the most common types of knowledge extracted by Relation Extraction systems (Mann, 2002), as the different Web People Search (WePS)[1] workshops show (Artiles et al., 2007). In this regard, the extraction of biographical information is a task whose performance may be improved if performed after a coreference solver, which previously established identity links between mentions of person entities, which are the main argument of a biographical relation (Suchanek et al., 2006).[2]

In order to take advantage of the benefits of coreference resolution, resources such as annotated corpora are needed. They are useful both for better understanding the behavior of this phenomenon and also for evaluating how coreference resolution systems perform.

Taking the above into account, this article presents three freely available corpora annotated with coreference links of person entities in Portuguese (pt), Galician (gl), and Spanish (sp). The corpora were designed with the two mentioned objectives: (i) providing new resources for researching how coreference behaves in these languages and (ii) evaluating how coreference resolution systems work with person entities. Furthermore, the presented resources are freely available in different formats, so they can be enlarged and improved collaboratively.[3]

Some related work in presented in Section 2. The main properties of the corpora as well as some distributional statistics are presented in Section 3. Then, the importance of coreference resolution for Information Extraction is shown in Section 4. Some final remarks are put forward in Section 5.

2. Related Work

The interest in having corpora with coreferential annotation was visible in the Message Understanding Conferences (MUC) (Fisher et al., 1995; Chinchor and Hirschmann, 1997), which started to develop guidelines and to build corpora for English with this kind of information.

Other evaluations such as the Anaphora Resolution Exercise (ARE) focused their attention on pronominal anaphora resolution and in nominal phrase coreference (Orasan et al., 2008). Previous works such as Mitkov et al. (2000) had continued developing annotation tools and corpora.

Based on the annotation of MUC schemes, Hoste (2005) proposed a coreference annotation scheme for Dutch, followed by the COREA Project (Bouma et al., 2007).

Recasens and Martí (2010) defined new annotation guidelines for coreference in Spanish and Catalan, by excluding some relations previously considered, such as part-whole coreference, bound anaphora or bridging reference.

This work also released corpora with coreferential annotation, and inspired the SemEval-2010 Task #1: Coreference Resolution in Multiple Languages (Recasens et al., 2010). Apart from Spanish and Catalan, this evaluation also

[1] http://nlp.uned.es/weps/weps-3

[2] In this paper, a *mention* is every instance of reference to a person, while an *entity* is the group of all the mentions referring to the same person in the text.

[3] http://gramatica.usc.es/~marcos/lrec.tar.bz2

Language	Text	Documents	Tokens
Portuguese	Journal	91	34k
	Wikipedia	6	17k
Galician	Journal	28	17k
	Wikipedia	29	25k
Spanish	Journal	27	18k
	Wikipedia	12	28k
Total	Journal	146	70k
	Wikipedia	47	71k
	Total	193	141k

Table 1: Size of the corpora in number of documents and tokens per language and text typology.

made available corpora for other languages such as English, Dutch, German and Italian.[4]

For Portuguese, Collovini et al. (2007) published Summ-it, a Brazilian Portuguese corpus focused on automatic summarization, which followed the MUC guidelines for including coreference annotation.

Finally, to the best of our knowledge, there is no corpus for Galician with any kind of coreference information.

Due to the lack of resources for Galician and Portuguese, this paper releases coreferentially annotated corpora with similar properties for these languages (and also for Spanish), which allow researchers from different fields to analyze this phenomenon and to evaluate coreference resolution systems.

3. The Corpora

The source texts used for building the corpora were collected from the Internet, in 2012, trying to cover different textual typologies as well as various linguistic varieties. Among the first ones, they contain journalistic and encyclopedic (Wikipedia) texts. Concerning the linguistic varieties, the corpora include texts of Portuguese of Portugal, Brazil, Mozambique and Angola, Galician and two varieties of Spanish: of Argentina and of Spain. Moreover, Wikipedia articles may belong to other varieties of Spanish and Portuguese.

Table 1 shows the size of the three presented corpora in number of documents and tokens, taking into account their text typology and language.[5]

Each corpus (Pt, Gl and Sp) has between 43k and 51k tokens (a total of $\approx 142k$ tokens). As the corpora are mainly focused on person entities, the distribution between journalistic and Wikipedia size is $\approx 35\%/65\%$, except for Portuguese, because of the interest in working with different varieties of this language, in a stage where the spelling system of Portuguese is changing due to an Orthographic Agreement.

In order to build the corpora, we first randomly selected journalistic news and encyclopedic articles (of people) from different Internet sources (taking into account their linguistic variety). These texts were tokenized, lemmatized and PoS-tagged with FreeLing (Padró and Stanilovsky, 2012). FreeLing was also used for performing Named Entity Recognition (NER) in Spanish, while Galician and Portuguese NER was carried out with other open-source tools (Garcia and Gamallo, 2010; Gamallo and Garcia, 2011; Garcia et al., 2012). Then, DepPattern was used for enriching the corpora with syntactic dependencies (Gamallo and González López, 2011). Finally, the coreferential annotation was manually added by two linguists following the SemEval-2010 Task #1 format (Recasens et al., 2010).

3.1. Annotation guidelines

Different expressions referring to the same discourse entity were annotated as coreferent when standing in a identity of referent relation. Moreover, predicative, appositive, and parenthetical expressions were also marked, even though these are not considered as coreferent expressions by some authors (Recasens and Martí, 2010). Although some entities appeared in different articles (and in different languages), the coreference annotation was performed using a single-document approach (cross-document annotation is planned for further versions of the corpora).

First column of Table 2 shows the linguistic units considered as markables during the coreferential annotation. Here, Nominal Phrases (NPs) with just a single token (e.g. a pronoun) or a compound proper noun, are classified using their PoS-tag (Personal Pronoun, Personal Name, etc.) instead of their syntactic label (Nominal Phrase). This way, in Table 2 Personal Names refers to every proper noun of a person occupying a full Nominal Phrase (NP), without specifiers. Therefore, Nominal Phrases only include NPs with at least two linguistic units: a head and a specifier ("Ayrton Senna" is classified as Personal Name, while "the driver Ayrton Senna" appears as a Nominal Phrase).

The annotation was carried out only when the mentions referred to personal entities identified in any part of the text (i.e., with at least one occurrence of a personal name).

Apart from the examples in Table 2, the corpora also include annotation of quoted speech. Thus, anaphoric personal pronouns occurring in quotes are linked to the entities they refer to: "Senna disse [...]: "**Eu** acho que"..." (*Senna said [...]: "**I** think that"...*).

3.2. Statistics

Concerning the distributional aspects of coreference, several statistics were computed for each corpus. Table 3 contains the percentage of each type of mention in the corpora, showing that personal names (as full NPs) and elliptical (zero) pronouns are the most frequent expressions for referring to person entities, with about 30% of mentions each. Moreover, nominal phrases and possessives (with 14% and 11%, respectively) also appear more than 10%.

On the contrary, the frequency of demonstrative and indefinite pronouns is scarce, with average values between 0.1 and 0.2 percent, respectively.

Table 3 also indicates that the distribution of the coreference links in the three analyzed languages is similar. Elliptical pronouns are less frequent in Portuguese than in Span-

[4] http://stel.ub.edu/semeval2010-coref/

[5] The statistics have been computed using the version 0.2 of the corpora. Further revisions might involve variations in these results.

	Port.		Galician		Spanish		Total		
	Jour.	Wiki	Jour.	Wiki	Jour.	Wiki	Jour.	Wiki	Total
Pers. Name	31.4	34.5	28.0	30.9	24.0	26.7	28.2	30.1	29.3
N. Phrase	24.3	11.5	21.0	8.5	12.8	11.9	19.7	10.5	14.4
Zero Pro.	26.6	26.0	30.7	36.2	34.0	32.7	29.9	32.5	31.4
Clitic Pro.	3.7	6.9	6.5	8.6	13.4	8.9	7.7	8.3	8.0
Relative Pro.	3.6	1.7	3.4	2.0	2.6	2.3	3.2	2.0	2.5
Personal Pro.	4.1	8.1	2.2	0.7	2.3	2.5	3.1	3.1	3.1
Demonst. Pro.	0.1	0	0	0.2	0.1	0.2	0.1	0.2	0.1
Indefinite Pro.	0.4	0.4	0.2	0.1	0.2	0.2	0.3	0.2	0.2
Possessive	5.8	11.1	8.1	12.9	10.5	14.5	7.9	13.5	10.9
Total mentions	2,418	1,561	925	2,631	1,826	2,634	5,169	6,826	11,995

Table 3: Distribution and total number of mentions per type, language and text typology.

Linguistic Unit	Text Example
Pers. Name	"Ayrton Senna cursou o primário nos Colégios Santana..."
	Ayrton Senna attended Colégio Santana...
Nom. Phrase	"Uma semana depois, o piloto brasileiro não conseguiu tempo..."
	A week later, the Brazilian driver did not get time...
Zero Pro.	"∅ Começou a competir oficialmente nas provas de kart..."
	[Ayrton Senna] Started to compete oficially in kart races...
Clitic Pro.	"Isso o deixou empatado com Nigel Mansell"
	This left him tied with Nigel Mansell
Relative Pro.	"[...] vinte pontos de diferença para Senna, que estava com zero"
	[...] twenty points of difference for Senna, who was with zero
Personal Pro.	"Ele sentia-se frustrado por..."
	He was frustrated by...
Demons. Pro.	"É bem provável que essa seja a esposa escolhida"
	It is quite probable that this is selected wife
Indefinite Pro.	"Ambos chegaram à F1..."
	Both of them arrived in F1
Possessive	"Senna começou sua carreira competindo..."
	Senna started his career competing...

Table 2: Coreference markables and examples (in Brazilian Portuguese and English)

ish (both in journalistic and in encyclopedic texts). On the other hand, NPs are less used for referring person entities in Spanish than in Portuguese. In average, Galician data have intermediate values between the two other languages.

Concerning the size of the entities (the number of mentions of each entity in the corpora), Table 4 includes the distribution of the person entities annotated (top rows). Entities with just one mention (*singletons*, which do not have coreferent expressions) are the most frequent ones (40%), while about 30% of them have more than three mentions.

Furthermore, central rows of the same table shows the average entity size as well as the size of the largest entity of each sub-corpus. Values have been computed for all the mentions and only for nominal ones (nominal phrases and personal names). The average size of the entities is similar in all the corpora, with values around 3.1% and 7.4% for all the mentions and only for the nominal ones, respectively.

About the largest entity of each sub-corpus, it is worth noting the difference between journalistic and Wikipedia texts, which have as a topic a specific person. Thus, some Portuguese and Spanish Wikipedia texts have entities with more than 600 mentions, while journal articles do not exceed 200 mentions. In Galician, the numbers in Wikipedia are noticeably lower due to the size of the articles, which are often smaller (Table 1).

Related to that, the bottom row of Table 4 shows the micro-average percentage of the mentions of the most frequent entity in each text, which represents the main person the text speaks of. Values from journalistic data are lower than the encyclopedic ones, since these articles are often focused on just one person. Notice that, in average, more than 60% mentions stands for a single entity in each text. It follows that most journalistic and encyclopedic documents speak of subjects whose content is organized around a pivot entity.

Some of these results differ from data extracted from other corpora with annotation of other types of entities. In this respect, Márquez et al. (2013) refers 86% of singletons in English and about 75% and 70% in Spanish and Catalan, respectively.

3.3. Format

Corpora are distributed in two different formats: (i) the default one, inspired in SemEval-2010 Coreference resolution task (Recasens et al., 2010); and (ii) in *brat* format,[6] an open-source tool which allows the annotator to visualize the text in a more efficient way.

[6] http://brat.nlplab.org/

	Portuguese		Galician		Spanish		Total		
	Journal	Wiki	Journal	Wiki	Journal	Wiki	Journal	Wiki	Total
1 Mentions	30.6	50.3	16.0	52.3	35.9	44.9	29.5	49.3	40.3
2 Mentions	17.0	17.3	24.8	18.4	21.5	18.0	19.5	18.0	18.7
3 Mentions	11.5	9.6	11.2	7.1	8.2	16.7	10.6	11.0	10.8
> 3 Mentions	40.1	22.8	48.0	22.2	34.4	20.5	40.4	21.8	30.2
Entity size	3.1	3.6	3.0	2.6	3.3	3.1	3.1	3.1	3.1
Entity size (Nom.)	5.6	7.9	6.7	6.6	9.1	8.3	7.2	7.6	7.4
Largest entity	35	218	20	121	26	146			
Largest entity (Nom.)	145	674	57	273	173	651			
Most frequent	50.2	62.5	53.4	69.9	59.3	64.5	54	66.1	60.1

Table 4: Distribution of entities according to their number of mentions in the corpora (top). Average size of the entities and size (in number of mentions) of the largest entity per language (center): values for all the entities and for nominal (Nom., including only NPs and proper nouns) ones. Distribution of the most frequent entity of each text (bottom).

token	...	*PoS*	...	*coref*
The	...	DET	...	(1
son	...	NOUN	...	_
of	...	PRP	...	_
the	...	DET	...	(2
Brazilian	...	ADJ	...	_
driver	...	NOUN	...	1)|2)
did	...	VERB	...	_

Figure 1: Example of a coreference annotation of the NPs "The son of the Brazilian driver".

The default format contains eleven columns with the following information: (1) Number of the token in the sentence, (2) Token, (3) Lemma, (4) FreeLing PoS-tag, (5) Basic PoS-tag, (6) Gender, (7) Number, (8) Syntactic head, (9) Syntactic label, (10) Named entity classification, (11) Coreference type and (12) coreference annotation.

Coreference annotation contains, for each mention, the *id* of the entity it belongs to as well as the *start* and *end* position of the mention, indicated with opening and closing parentheses (Figure 1). When a token belongs to more than one mention, '|' is used for separating the entities *ids*.

The corpora distributed in SemEval-2010 contain more linguistic information than the presented in this paper, such as semantic role labelling, and the PoS-tags as well as the syntactic labels slightly differ. Apart from that, the main difference in the annotation between those corpora is related to the punctuation. Entities in SemEval-2010 data include the punctuation around them, while in this corpora only the lexical tokens are considered part of the entity (except in entities which contain punctuation inside them).

4. Coreference Resolution and Information Extraction

This section presents the results of a simple evaluation performed to know how the coreference solving of person entities in a corpus improves the performance of Information Extraction systems.

In order to perform the test, a multilingual Open Information Extraction system, called *DepOE* (Gamallo et al., 2012), was used to extract verb based triples from the three presented corpora.

The test measured two different aspects: First, the increase in recall caused by the pronominal coreference solving. Second, the enrichment of the extraction due to the resolution of nominal coreference.

The impact of coreference resolution was calculated as follows: First, *DepOE* was applied on the three presented corpora. For measuring the recall increase, the occurrences of personal and elliptical pronouns as the first argument of the extracted triples were counted and divided by the number of personal names occupying the same position.

Concerning the enrichment of the nominal extractions, the improvement was measured by dividing the number of person entities in a text by the number of different nominal mentions in the same text. Thus, this value suggests how richer the extraction is, taking into account that different mentions from the same entity are not clustered if the coreference was not previously solved.

	Port.	Galician	Spanish
Pronominal Coref.	2.71	3.52	4.3
Nominal Coref.	48%	59%	55%

Table 5: Pronominal Coref. indicates the recall increase. Nominal Coref. is the estimation of the extraction enrichment.

First row in Table 5 shows the recall increase in the three corpora, while the second row indicates the enrichment of the extraction due to the solving of nominal coreference.

Information Extraction systems obtain structured information from free text, but some of the obtained knowledge is not very useful, since their main argument would be a pronoun (or other linguistic unit) without reference.

Furthermore, many other extractions are not correctly classified due to the lack of information concerning the entities they refer to, involving incoherence in the extraction.

Solving pronominal coreference may increase the recall between 2.7 and 4.3 times. Moreover, correctly clustering the nominal mentions leads to an improvement of the extraction in more than 50% in average.

5. Conclusions

This paper presents three different corpora with manual annotation of coreferent personal entities in three languages: Portuguese, Galician and Spanish. Taking into account the importance of coreference both in general linguistics and in Natural Language Processing, this paper provides useful resources for its study and for evaluating coreference resolution systems in different languages. Furthermore, their freely distribution allows other contributors to use, enlarge, or modify these corpora. The paper also includes a statistical study which shows several distributional aspects of coreference in person entities. Finally, an empirical evaluation of coreference resolution for Information Extraction was also performed, proving that accurately solving this phenomenon notoriously improves the performance of an Open Information Extraction system.

Acknowledgements

This work has been supported by Ministerio de Ciencia e Innovación, within the project OntoPedia, reference: FFI2010-14986, and by the Galician Government, within the project with reference: EM2013/041.

6. References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, Czech Republic. Association for Computational Linguistics.

Michel Banko and Oren Etzioni. 2008. The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, volume 8, pages 28–36. Association for Computational Linguistics.

Gosse Bouma, Walter Daelemans, Iris Hendrickx, Véronique Hoste, and Anne-Marie Mineur. 2007. The COREA-project. Manual for the annotation of coreference in Dutch texts.

Nance Chinchor and Lynette Hirschmann. 1997. MUC-7 Coreference Task Definition (Version 3.0). In *Proceedings of Message Understanding Conference 7*, volume 7. Association for Computational Linguistics.

Sandra Collovini, Thiago I. Carbonel, Juliana Fuchs Thiesen, Jorge C. Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visandoa sumarização automática.

Davir Fisher, Stepehn Soderland, Fangfang Feng, and Wendy Lehnert. 1995. Description of the UMass system as used for MUC-6. In *Proceedings of Message Understanding Conference 6*, pages 127–140. Association for Computational Linguistics.

Pablo Gamallo and Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. In *Progress in Artificial Intelligence (LNCS/LNAI)*, volume 7026/2011, pages 610–623. Springer-Verlag.

Pablo Gamallo and Isaac González López. 2011. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International Journal of Corpus Linguistics*, 16(1):45–71.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based Open Information Extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18. Association for Computational Linguistics.

Marcos Garcia and Pablo Gamallo. 2010. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2):59–67.

Marcos Garcia, Iria Gayo, and Isaac González López. 2012. Identificação e Classificação de Entidades Mencionadas em Galego. *Estudos de Lingüística Galega*, 4:13–25.

Peter C. Gordon and Randall Hendrick. 1998. The representation and processing of coreference in discourse. *Cognitive Science*, 22.4:389–424.

Véronique Hoste. 2005. *Optimization issues in machine learning of coreference resolution*. Ph.D. thesis.

Gideon S. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *Proceedings of the 2002 workshop on Building and using semantic networks*, volume 11. Association for Computational Linguistics.

Ruslan Mitkov, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones, and Violeta Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, pages 49–58.

Lluís Márquez, Marta Recasens, and Emili Sapena. 2013. Coreference resolution: an empirical study based on SemEval-2010 shared Task 1.

Constantin Orasan, Dan Cristea, Ruslan Mitkov, and António Branco. 2008. Anaphora Resolution Exercise: an Overview. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. ELRA.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Turkey. ELRA.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44.4:315–345.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pages 1–8. Association for Computational Linguistics.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Kael Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management: an International Journal*, 43(6):1663–1680.

Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. LEILA: Learning to extract information by linguistic analysis. In *Workshop on Ontology Population at ACL/COLING*. Association for Computational Linguistics.