Natural Language Engineering **21** (4): 519–551. © Cambridge University Press 2013
doi:10.1017/S1351324913000314

Exploring the effectiveness of linguistic knowledge for biographical relation extraction†

MARCOS GARCIA and PABLO GAMALLO
Centro Singular de Investigación en Tecnoloxías da Información (CITIUS), University of Santiago de Compostela, Coruña, Spain
e-mail: {marcos.garcia.gonzalez, pablo.gamallo} @usc.es

(*Received 19 June 2012; revised 16 September 2013; accepted 18 September 2013; first published online 18 October 2013*)

Abstract

Machine learning techniques have been implemented to extract instances of semantic relations using diverse features based on linguistic knowledge, such as tokens, lemmas, PoS-tags, or dependency paths. However, there has been little work aiming to know which of these features works better in the relation extraction task, and less in languages other than English. In this paper, various features representing different levels of linguistic knowledge are systematically evaluated for biographical relation extraction. The effectiveness of these features was measured by training several supervised classifiers that only differ in the type of linguistic knowledge used to define their features. The experiments performed in this paper show that some basic linguistic knowledge (provided by lemmas and their combination in bigrams) behaves better than other complex features, such as those based on syntactic analysis. Furthermore, some feature combinations using different levels of analysis are proposed in order (i) to avoid feature overlapping as well as (ii) to evaluate the use of computationally inexpensive and widespread tools such as tokenization and lemmatization. This paper also describes two new freely available corpora for biographical relation extraction in Portuguese and Spanish, built by means of a distant-supervision strategy. Experiments were performed with five semantic relations and two languages, using these corpora.

1 Introduction

Relation extraction (RE) is a subtask of information extraction whose aim is to automatically extract instances of semantically related entities from corpora.

Biographical relation extraction (BRE) attempts to obtain pairs belonging to a biographical relation, such as (PERSON)`hasBirthDate`(DATE), (PERSON)`hasDeathPlace` (LOCATION), or (PERSON)`hasProfession`(OCCUPATION).

† This work has been supported by Ministerio de Ciencia e Innovación, within the project OntoPedia, ref: FFI2010-14986.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

For example, from a sentence like '*Kimberley Deal (born June 10, 1961) is an American singer, songwriter and musician*,' the goal of a BRE system is to extract the following structured data:

- *Kimberley Deal* `hasBirthDate` *10 of June of 1961*
- *Kimberley Deal* `hasNationality` *American*
- *Kimberley Deal* `hasProfession` *singer*
- *Kimberley Deal* `hasProfession` *songwriter*
- *Kimberley Deal* `hasProfession` *musician*

These data can be arranged, for instance, into proper noun ontologies, which are useful for many applications such as question answering systems (Mann 2002; Ravichandran and Hovy 2002) or information retrieval engines (Wan *et al.* 2005).

In recent years, several strategies have been applied for RE, like semi-supervised classifiers or more unsupervised approaches. These methods are focused on minimizing the effort of manually annotating corpora. However, one of the main approaches for extracting high-quality instances of semantic relations from unstructured text is still the use of supervised classifiers, trained with annotated sentences containing pairs of the desired relations. These classifiers may contain different kinds of features: from orthographic and lexical data to more complex features such as part-of-speech (PoS) tags, semantic knowledge, or syntactic information.

In order to build accurate feature-based classifiers, some works adjust the feature space to a specific relation and/or pattern, creating high-quality extractors (Fleischman, Hovy and Echihabi 2003). Other works showed that several features (and combinations of them) behave well with some relations, while their performance decrease in other scenarios (Agichtein 2005). However, there has been little research in feature engineering for RE, and it has been usually performed only for English (Kambhatla 2004; Zhou *et al.* 2005).

In this respect, evaluating the effectiveness of different linguistic knowledge for RE is important not only for this task, but also for theoretical and computational linguistics in general, in order to know the best way for representing linguistic structures containing semantic relations.

The main objective of this paper is to perform an evaluation of different features for BRE for Portuguese (pt) and Spanish (sp). Additionally, two freely available corpora for this task were also created and analyzed.

Taking this into account, a systematic evaluation of several linguistic features is performed. Extraction is carried out between entities appearing in the same sentence, by means of generic features that have been used for RE, extracted from different levels of linguistic knowledge.

Concerning the corpora, two collections of sentences containing biographical relations (one for Portuguese and another one for Spanish) were created and annotated with the following semantic relations: `hasBirthPlace`, `hasBirthDate`, `hasDeathPlace`, `hasDeathDate` and `hasProfession`.[1] Furthermore, a careful analysis of the relation patterns was carried out, aimed at knowing how the target

[1] Corpora are available at `http://gramatica.usc.es/~marcos/corpora_nle.tgz`

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

relations are expressed in Portuguese and Spanish. The manual effort of building annotated corpora was reduced by applying a method based on distant-supervision (Mintz *et al*. 2009). Then, machine learning algorithms were applied in order to build and to evaluate several classifiers, differing only in the types of linguistic features they have.

The results of several experiments show that using lemmatization—as well as basic semantic knowledge provided by named entity recognizers—improves the accuracy of *bag-of-words* classifiers. Furthermore, pseudo-syntactic information (mainly represented by bigrams of lemmas) can be used to avoid computationally expensive processes such as syntactic parsing, which, in turn, seems not to provide great benefits on the system.

The remainder of this paper is organized as follows. Section 2 introduces some related work. Then, Section 3 provides a general view of how the corpora were constructed as well as an analysis of their linguistic structures. Next, Section 4 is focused on the description of the features, particularly on the types of linguistic knowledge used by the classifiers. Then, Section 5 reports the diverse experiments performed over the Portuguese and Spanish texts. Finally, some conclusions are put forward in Section 6.

2 Related work

There has been a large body of work on extraction of semantic knowledge from unstructured text in recent years. In this section, some seminal papers on RE are introduced, followed by several state-of-the-art approaches for this task. Then, various works focused on BRE are presented, paying special attention to the linguistic knowledge they use. After that, some papers that evaluate different features in RE are shown. Finally, several specific works of RE for Portuguese and Spanish are also introduced.

Hearst (1992) was the first one to experiment with a pattern-matching approach for the identification of semantic relations. She made use of a small set of initial patterns to get hypernymy relations, thus increasing the set of patterns with a bootstrapping technique. Brin (1998) applies a similar method, but performs an evaluation of the new patterns, in order to promote only those that show a good precision.

Other work that iteratively learns new extraction patterns is Agichtein and Gravano (2000), which present *Snowball*. This tool extracts and evaluates the quality of patterns in order to obtain new semantically related pairs without human intervention. Ravichandran and Hovy (2002) make use of question answering pair examples to automatically extract patterns. KnowItAll (Etzioni *et al*. 2004) also extracts instances of semantic relations in an unsupervised fashion, by learning new extraction rules in a bootstrapping process. *Espresso* is another tool for RE from the web. It starts by using patterns with high recall and lower precision. Then, *Espresso* uses the web in order to filter those patterns with low precision (Pantel and Pennacchiotti 2006). Bunescu and Mooney (2007) also use small sets of positive and negative seed pairs for obtaining sentences containing the desired relations.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Recently, new approaches for RE have been proposed. *Distant-supervision* takes advantage of large repositories of related pairs to automatically classify sentences, then used for training classifiers (Mintz *et al.* 2009).

Another paradigm, *Open Information Extraction*, attempts to extract large sets of related pairs without manually specifying semantic relations (Banko *et al.* 2007). Similarly, *woe* performs open information extraction taking advantage of the semi-structured resources of Wikipedia (Wu and Weld 2010). *Woe* uses syntactic information extracted from parsing Wikipedia, but does not outperform new approaches of open information extraction (Etzioni *et al.* 2011; Gamallo, Garcia and Fernández-Lanza 2012).

2.1 Biographical relation extraction

More related to our work, several papers and competitions focus on the extraction of pairs belonging to biographical relations.

Mann (2002) describes a method for creating ontologies of proper nouns for question answering systems, by selecting instances of a common noun followed immediately by a proper noun (the *composer*$_{\text{(common noun)}}$ *Wolfang Amadeus Mozart*$_{\text{(proper noun)}}$). Inspired by this work, Fleischman *et al.* (2003) apply machine learning algorithms in order to classify instances of similar 'common noun–proper noun' examples.

Jijkoun, De Rijke and Mur (2004) evaluate different surface and syntactic patterns for obtaining pairs for question answering. Their results show that syntactic parsing improves the performance of the question answering systems, despite the fact that the extraction is less precise than that obtained through surface patterns.

Garera and Yarowsky (2009) present several approaches for extracting biographical facts: their objective is to include structural information of biographical texts (position of the facts in the document, extraction of implicit knowledge, etc.) into the extractors, thus improving the performance of surface patterns for RE.

Liu *et al.* (2010) describe *BioSnowball*, a framework to collect personal information from the web in order to generate Wikipedia-style pages.

Other works explore the extraction of biographical knowledge, although they are not focused only on this kind of relations. LEILA (Suchanek, Ifrim and Weikum 2006) is a system that can extract instances of semantic relations from natural language documents, by using syntactic parsing. Finally, Pasca *et al.* (2006) also extract a large set of `hasBirthDate` pairs by generalizing patterns found with a small set of initial seed pairs.

In recent years, two competitions have included tasks about biographical fact extraction: the Knowledge Base Population (KBP) Slot Filling task (in the Text Analysis Conference, TAC)[2] and the Person Attribute Extraction (in the Web People Search Evaluation Campaign, WePS),[3] both of them since 2009. The objective of these competitions is to extract attribute values from a predefined set of biographical

[2] `http://www.nist.gov/tac/data/index.html`
[3] `http://nlp.uned.es/weps/`

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

relations. In some scenarios, the systems should perform clustering of web pages, coreference resolution, as well as cross-sentence and cross-document fact extraction, so the tasks are not directly comparable to the approach followed in this paper.

2.2 Linguistic knowledge

Different RE systems, from pattern-matching approaches to more complex statistical methods, use linguistic knowledge in order to improve the extraction of related pairs.

First, strategies for RE (Hearst 1992) simply applied lexical patterns ('`entity_1` *such as* `entity_2`') to obtain the related pairs. Then, in order to increase the recall of their systems, other works performed different types of pattern generalization, by means of the longest common substring (Finkelstein-Landau and Morin 1999), applying a suffix tree (Ravichandran and Hovy 2002), or by computing the edit distance between the different patterns (Ruiz-Casado, Alfonseca and Castells 2005).

The above mentioned Mann (2002) included PoS-tags in the patterns for extracting specific relations such as `hasProfession`. In a similar way, *BioSnowball* (Liu *et al.* 2010) also uses a combination of lexical and PoS-tag sequences for representing the patterns.

A novelty in the work of Suchanek *et al.* (2006) was the use of syntactic analysis (provided by a link grammar) in the system. Similarly, Akbik and Broß (2009) apply linguistic patterns (using dependency grammars) for populating a semantic wiki from the English Wikipedia.

Some other works also use syntactic parsing for RE. Bunescu and Mooney (2005), Snow, Jurafsky and Ng (2005), and Nguyen, Matsuo and Ishizuka (2007) created different statistical models using the results of syntactic parsing. More recently, Yan *et al.* (2009) perform RE by combining surface patterns extracted from the Web with dependency patterns obtained from the parsing of Wikipedia.

RGAI (Nagy and Farkas 2010) was the system with the best results in the Attribute Extraction subtask of WePS-3. Its method consists of two parts: first, candidate paragraphs are extracted if they contain relevant lexical information for each relation (e.g., *born*, *birth* or *birthday* for `hasBirthDate`); then, a validator classifies the candidate attribute as positive or negative based on regular expressions, string-matching, and other relation-dependent heuristics.

Sun *et al.* (2011) achieved the top results at the 2011 KBP Slot Filling task, by training distant-supervision classifiers. Concerning the linguistic knowledge they use, this system obtains features from the token sequence (words after and before the candidate entities, intermediate pattern of tokens), from the semantic classification of the entities (PERSON, ORGANIZATION, etc.), and from the dependency tree (the shortest dependency path between the candidate pair, and the head and dependent word of each candidate entity).

Concerning the effectiveness of various linguistic features, some works evaluated the impact of different knowledge over the ACE corpus (for generic RE). Kambhatla (2004) and Zhou *et al.* (2005) show that the effective incorporation of different features based on linguistic knowledge improves the performance of the classifiers.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

However, Jiang and Zhai (2007) conclude that adding more complex features may hurt the performance of the extractors.

Other strategies using kernel methods (as well as several kernel combinations) were also explored in order to improve the RE using the ACE corpus (Zhao and Grishman 2005; Zhang *et al.* 2006; Zhou *et al.* 2007; Nguyen, Moschitti and Riccardi 2009).

2.3 Relation extraction for Portuguese and Spanish

As it was shown in the previous sections, most of the work on RE is performed for English. However, there are some papers on RE for Portuguese and Spanish that need to be mentioned.

Specifically for Portuguese, several works were focused on different types of RE. Oliveira *et al.* (2008) deal with the extraction of relations between concepts—such as `hypernym`, `hyponym`, `part_of`, `inclusion`, etc.—from a Portuguese dictionary. Using this strategy as well as other extractions based on pattern-matching, the Onto.PT project aims to automatically construct a lexical ontology for Portuguese (Oliveira and Gomes 2010). Other works such as (Costa and Branco 2012) identify events and temporal expressions in unstructured text.

The second HAREM evaluation included a shared task on generic RE between named entities (ReRelEM) (Mota and Santos 2008). Three different systems were presented: REMBRANDT uses knowledge extracted from Wikipedia for Named Entity Recognition (NER), and a set of heuristics and grammatical rules for identifying the relations (Cardoso 2008). SEI-Geo applies *Hearst patterns* (and other strategies such as using trigger words and verbs) for establishing relations between locations (Chaves 2008). Finally, SeRELeP also attempts to identify relations by applying simple rules on previously identified named entities (Bruckschen *et al.* 2008). Concerning biographical relations, Soares, Martins and Calado (2011) evaluate various algorithms and features for classifying sentences (but not pairs of entities) containing this kind of relations.

For Spanish, Sánchez-Cuadrado *et al.* (2003) take advantage of syntactico-semantic patterns for building a thesaurus of zoology. Sierra *et al.* (2008) propose the use of verbal patterns for extracting (i) `hypernym` and `hyponym` relations, (ii) `synonym`, as well as (iii) `individuation` (`quantity/mass` and `member/group`) instances. Other works such as Soler-Alcina (2008) also apply lexical patterns for obtaining instances of the `part-whole` relation (in the domain of ceramics).

Aguado de Cea *et al.* (2008) present a method (and a tool) for re-using ontology design patterns (by associating them to lexico-syntactic patterns) for ontology population (for English, German, and Spanish).

Garcia and Gamallo (2011a) compare the impact of several features for distant-supervision extraction in Spanish, focusing on the `hasProfession` relation.

Finally, Garcia and Gamallo (2011b) apply a sentence compression method in order to increase the coverage of a RE system in Spanish and Portuguese.

In this paper, the effectiveness of several types of linguistic features for BRE is explored, by training different feature-based classifiers: first, by using simple lexical

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

features, and then by evaluating the incremental addition of more complex features based on morphosyntactic analysis, pseudo-syntactic representation, as well as on syntactic knowledge.

3 The corpora

In this section, the method for obtaining the corpora used during the experiments as well as their main characteristics are presented. Two corpora were built, one for Portuguese and another one for Spanish.

3.1 Building the corpora

Manually creating a labeled corpus is an expensive process, but it is crucial for many applications relying on high-quality data for building statistical models. So, in order to speed up this process, a strategy inspired by distant-supervision was applied.

Distant-supervision assumes that if two entities participate in a relation (e.g., `hasBirthPlace`: 'Woody Allen', 'The Bronx'), any sentence containing these two entities might express the same relation (`hasBirthPlace`) (Mintz *et al.* 2009). Thus, obtaining large sets of pairs belonging to specific relations (from databases such as Freebase, DBpedia, etc.) permits to automatically classify extracted sentences containing different pairs as positive or negative for the desired relations.

Although this strategy gave good results for some relations and datasets (Hoffmann, Zhang and Weld 2010), Riedel, Yao and McCallum (2010) observed that it might produce a large amount of noisy patterns if the knowledge base and the text do not belong to the same—or similar—domain. Taking this into account, the corpora used in this paper were created by manually correcting the output of a distant-supervision extraction.

It is worth mentioning that applying this method for languages other than English is not always feasible, since the attributes of many relations are language-specific, and the most common (and largest) knowledge bases are mainly focused on English.[4]

So, language-independent pairs (containing dates) were obtained from Freebase and the English DBpedia. The other sets of pairs (language-dependent) were extracted from the Portuguese and Spanish Wikipedia infoboxes (and also from the incipient versions of DBpedia for these languages).

The following strategy was applied:

(1) First, large sets of pairs for each relation and language were automatically extracted from the mentioned sources. For `hasBirthDate` and `hasDeathDate`, 460,703 pairs were used. For `hasBirthPlace`, 45,588 and 8,952 for Portuguese and Spanish, respectively. The amount of `hasDeathPlace` pairs was 11,664 (pt) and 1,319 (sp). For `hasProfession`, only the most common occupation nouns were used, by selecting those with more than 20 occurrences in the seed

[4] As an example, at the time of building the corpora, DBpedia for English contained 433,042 instances of `hasBirthDate`, while the Portuguese and Spanish versions had 47,460 and 1,243, respectively.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

list, in order to avoid noisy sentences. Therefore, 68 and 96 different nouns instantiating this relation were used for Portuguese and Spanish, respectively.

(2) Then, these pairs were used for extracting sentences containing known entities from the Portuguese and Spanish versions of Wikipedia (with 700 MB and 1.6 GB of data, respectively). Furthermore, 1 GB and 225 MB of journalistic corpora were also used for Portuguese (*Público*) and for Spanish (*El País*). The final extraction had Wikipedia/newspaper ratios of 90%/10% (pt) and 93%/7% (sp).

(3) Both Portuguese and Spanish corpora were first analyzed by some modules (sentence chunking, tokenizing, lemmatizing, PoS-tagging, and dates and named entity recognition) of FreeLing (Padró *et al.* 2010). For avoiding noisy sentences, they were selected only when the entities they contained were exactly the same as in the seed pairs. Moreover, these entities had been classified by FreeLing with the expected class (e.g., PERSON, LOCATION), except for dates, which were selected if at least the year was the same as in the seed examples.

(4) The extracted sentences were automatically labeled as positive for a specific relation if they contain both entities of a pair. Otherwise, they were marked as negative.

(5) Sentences were enriched with syntactic knowledge in order to build a *linguistic structure* (explained below), useful for extracting different types of features that will be evaluated later. DepPattern, a multilingual suite for syntactic analysis (Gamallo and González 2011) was applied for generating dependency graphs from the sentences.

(6) Then, the label (positive or negative) of each candidate pair as well as the boundary and the semantic classification (PERSON, LOCATION, DATE, etc.) of the two entities were manually corrected.

(7) Finally, the entities of each related pair were substituted by **X** (PERSON) and **Y** (attribute: **Y_Loc**, **Y_Dat** or **Y_Occ**).

Both corpora for Portuguese and Spanish contain five biographical relations: `hasBirthPlace`, `hasBirthDate`, `hasDeathPlace`, `hasDeathDate`, and `hasProfession`. Moreover, the corpora have 268,469, and 152,817 tokens for Portuguese and Spanish, respectively.

Table 1 shows the number of labeled pairs for each relation and language. Positive pairs are used as negative examples for different relations with the same entity classes. Thus, a sentence such as 'PERSON was born in LOCATION' (positive for `hasBirthPlace`) will be used as a negative example for `hasDeathPlace`.[5]

3.2 Linguistic structure

Each labeled, tagged, and parsed sentence represents a linguistic structure containing all the relevant information required by the RE system. A linguistic structure or parsed sentence can be conceived as a knowledge-rich space incorporating several levels of linguistic information.

[5] Apart from minor exceptions, not taken into account, where a sentence can be positive for more than one relation: 'PERSON was born and died in LOCATION.'

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Table 1. *Number of candidate pairs for each relation and language of the annotated corpora.* Pos. *stands for positive pairs for the semantic relation and* Neg. *for the negative ones*

	Portuguese		Spanish	
Relation	Pos.	Neg.	Pos.	Neg.
`hasBirthPlace`	1,312	1,186	493	1,149
`hasBirthDate`	552	510	419	662
`hasDeathPlace`	879	5,040	563	833
`hasDeathDate`	402	835	570	771
`hasProfession`	1,277	740	1,595	715
Total	4,422	8,311	3,640	4,130

Sentence: Kimberley Deal (born June 10, 1961) is an American singer.
Relation: Kimberley_Deal `hasProfession` singer
Polarity: positive
Linguistic Structure:

pos	token	lemma	tag	head	label
0	**X**	Kimberly_Deal	PERSON	5	Subj
1	(	(	FA	2	Punct
2	born	born	VB	0	Ad
3	June_10_,_1961	06/10/1961	DATE	2	Obj
4	)	)	FC	2	Punct
5	is	be	VB	9	ROOT
6	an	a	DI	8	Spec
7	American	american	ADJ	8	Modif
8	**Y_Occ**	singer	NOUN	5	Attrib
9	.	.	FT	-	-

Fig. 1. Example of a sentence (in English) labeled for `hasProfession` with the related entities (**X** and **Y_Occ**), the polarity, and its linguistic structure.

Figure 1 contains an example of the `hasProfession` relation, where the original sentence, the related terms with a polarity label, and the linguistic structure are shown. Note that the same sentence will be also labeled as positive for `hasBirthDate`, with the attribute **Y_Dat** being the entity in *pos* 3.[6]

In the example (Figure 1), columns 1, 2, 3, and 4 respectively stand for position, token, lemma, and PoS-tag. Last two columns contain syntactic information: 5 identifies the head position of the current token, while the label of the dependency

[6] The meaning of the PoS-tags are as follows: FA, open parenthesis; VB, verb; FC, close parenthesis; DI, indefinite determiner; ADJ, adjective; NOUN, common noun; FT, period. Dependency labels mean: Subj, subject; Punct, punctuation; Ad, adjunct; Obj, object; Spec, specifier; Modif, modifier; Attrib, attribute.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

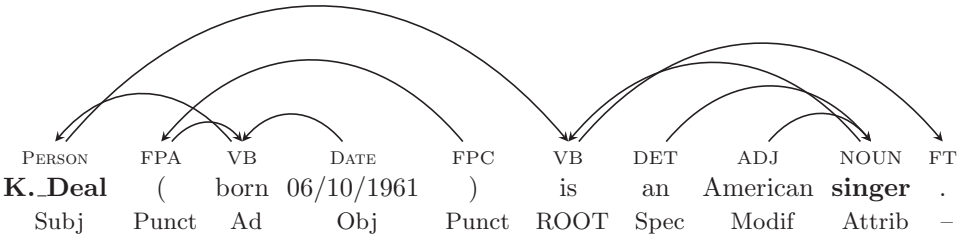


Fig. 2. Example of a dependency graph. First row contains the PoS-tag, enriched with semantic classes, while the syntactic function is shown in the third row.

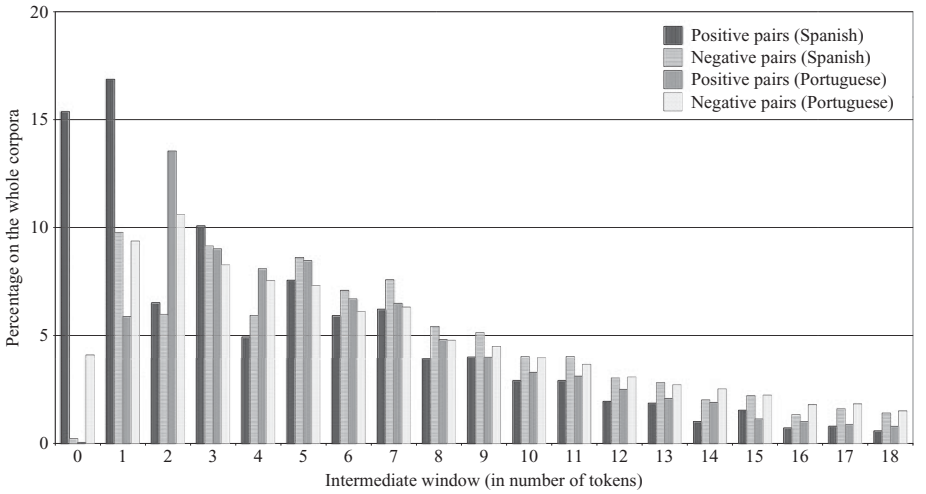


Fig. 3. Histogram of positive/negative pairs *versus* intermediate context in number of tokens. Values are micro-average results of the five analyzed relations.

is shown by column 6. This structure was inspired by the output parsing format defined in Lin (2003), adopted by the evaluation of the CoNLL Shared Tasks.

The syntactic analysis was obtained by a dependency parser. Dependency parsing establishes binary links between a head and a dependent, generating a dependency graph from a sentence. Dependency links encode information about the head and the dependent of a relation, as well as the syntactic function of this dependency. Figure 2 exemplifies a dependency graph of the previous example sentence, where the arrows point from the dependent to the head.

3.3 Linguistic properties of the corpora

Before extracting the features for building the classifiers, an analysis of the patterns containing the desired relations was carefully performed in the corpora: first, by creating a histogram of the number of tokens that occur between the two candidate entities, and second, by finding the most common patterns for each relation and language.

Figure 3 shows a histogram that summarizes the intermediate number of tokens (from 0 to 18) between the candidate entities in the Portuguese and Spanish corpora.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Table 2. *Best patterns (taking into account their precision and their percentage in the sets of positive sentences) per language and relation. Elements in curly brackets are—each one—optional, vertical bars represent disjunction, and* NP *PoS-tags refer to a proper noun. For space reasons, patterns are highly simplified by omitting some verbal inflections and other optional elements.* BD *means* `hasBirthDate`*;* BP, `hasBirthPlace`*;* DD, `hasDeathDate`*;* DP, `hasDeathPlace`*, and* Pr. *refers to* `hasProfession`*.* % *is the percentage of each pattern in the positive sentences of each relation.* Prec *is its precision in the corpus*

Lang.	Rel.	Pattern	%	Prec
Pt	BD	**X (** {NP ,|; NP ,|;} **Y_Dat**	30.4	96
		X {(|,} **nasceu|nascido** {em NP ,} **em** {o}|**a Y_Dat**	13.4	97
	BP	**X** {,} **nasceu|nascido** {em a|o NP ,} **em o Y_Loc**	49.3	98
		X {,} **(** {Dat , NP , em} **Y_Loc**	14.2	94
	DD	**X (** {NP , NP ,} **Dat** {-|;|,} {NP ,} **Y_Dat**	20.7	100
		X {,} **morreu|faleceu em Y_Dat**	2.5	100
	DP	**X** {,} **(NP ,** {NP ,} **Dat** {-|; Dat} **Y_Loc**	21.1	99
		X {,} **morreu|faleceu** {em Dat} **em** {o|a} **Y_Loc**	2.2	100
	Pr.	**Y_Occ** {,} {de a NP} **X**	27.1	99
		X ,|é|foi {um|uma} **Y_Occ**	13.5	97
Sp	BD	**X (** {NP ,|; NP ,|} **Y_Dat**	45.8	94
		X {(|,} **nació|nacido** {en NP} **el Y_Dat**	6.4	100
	BP	**X** {,} **(** {NP|Dat ,} **Y_Loc**	39.8	98
		X nació|nacido {el Dat ,} **en Y_Loc**	9.3	100
	DD	**X (** {NP , NP ,} **Dat** {-|;|, NP ,} **Y_Dat**	24.2	95
		X falleció|murió {en NP} **en Y_Dat**	3.5	95
	DP	**X** {,} **(NP ,** {NP ,} **Dat** {-|;|, Dat ,} **Y_Loc**	34.0	100
		X {,} **falleció|murió** {en Dat} **en Y_Loc**	1.2	100
	Pr.	**Y_Occ** {de NP ,} **X**	36.0	99
		X ,|es|fue {un|una} **Y_Occ**	5.2	96

Even though the intermediate number of tokens differs for each relation (as the most common patterns will show), the histogram of Figure 3 clearly proves that semantic relations occur more frequently the closer the entities are (with the best ratio with less than eight intermediate tokens, depending on the language and on the relation). As the size of this window increases, more negative pairs appear, and little positive ones are found.

Concerning the patterns that express biographical relations, Table 2 includes the two best patterns per relation and language, taking into account their percentage in the sets of positive sentences as well as their precision. Note that these patterns are greatly simplified by using disjunctions as well as by omitting some optional elements.

The taxonomy presented in Table 2 shows that target relations often occur in biographical-specific patterns with high precision, which contain in parenthesis biographical facts about a previously mentioned person. This table also indicates that some relations heavily depend on these patterns (which represent between 36% and 49% of the positive pairs), while others occur in a wide range of structures.

However, it is worth mentioning that, except for these frequent and high-precision patterns, other structures are very ambiguous and their use to identify a specific relation may depend on lexical data outside the intermediate window.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

In order to represent the patterns, these analyses point out the importance of choosing the best intermediate window as well as the right features. Note that the patterns underlying biographical sentences not only are based on lexical and syntactic structures, but also contain a great variety of punctuation marks designing relevant linguistic information.

4 Features

In order to evaluate the effectiveness of the linguistic knowledge for RE, features were organized from shallow to deep analysis, according to the linguistic complexity they contain. Therefore, this section presents and discusses the different features used for training the classifiers.

4.1 Basic features

Before introducing the basic features, it is worth noting that the selection of candidate entities is restricted to specific semantic classes for each relation. Thus, **X** will always be a PERSON proper noun, and **Y** will be DATE in `hasBirthDate` and `hasDeathDate` and LOCATION in `hasBirthPlace` and `hasDeathPlace`. Finally, **Y** will be an occupation noun in `hasProfession`. The effectiveness of this restriction will be further analyzed by using a baseline that does not take advantage of this semantic classification.

Apart from the use of NER tools, the first category of features is not purely linguistic. It includes information about the position of both **X** and **Y** in the sentence. This category contains four different features (their values come from the example shown in Figure 1):

- Absolute position of **X** in the sentence: 0
- Absolute position of **Y** in the sentence: 8
- Direction of the relation: **X_Y** (1) or **Y_X** (2): 1
- Distance (in number of tokens) between both entities: 7

Although these features do not characterize the semantic relation at all, they are widely used in the literature, and preliminary experiments showed they are helpful for combining with some linguistic features (explained below) that do not include implicit information about the position of the entities. Thus, these features will be combined with the subsequent categories including lexical, morphosyntactic, pseudo-syntactic, and syntactic knowledge.

4.2 Lexical

These features concern the use of the lexical units present in the linguistic structure. Two main types of features were taken into account:

- `Tok`: including the tokens between the two related entities as well as some surrounding elements (defined below)
- `Lem`: using lemmas, the normalized form of each token

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Tokens are common features used to characterize linguistic structures containing semantic relations. Using them as *bag-of-words* adds important lexical information. An example of the `Tok` features extracted from the example can be seen below:

Sentence: **X** (born June 10, 1961) is an American **Y_Occ**
`Tok`: <(>, <born>, <DATE>, <)>, <is>, <an>, <American>

Note that `Tok` and `Lem` are not simple *bag-of-words* features. They include the boundaries of complex entities such as compound proper nouns (e.g., *Kimberly_Deal*) and temporal expressions (e.g., *June 10, 1961*), by means of different modules included in FreeLing.

In highly inflected languages such as Portuguese and Spanish, using tokens may involve data sparseness. For example, each verb has about 50 different forms in these languages. Therefore, using lemmas as features allows the classifier to generalize this information. The previous example sentence will be encoded as follows:

`Lem`: <(>, <bear>, <DATE>, <)>, <be>, <a>, <american>

Note that these features will also be extracted from similar sentences such as '**X** (born June 10, 1961) **was** an American **Y**'. So, `Lem` features reduce the data sparseness even if they lose some morphosyntactic information included in the inflection of nominal and verbal forms.

Using *bag-of-words* strategies (with tokens, lemmas, and other features) for representing patterns involves the definition of two thresholds. First, the surrounding context of **X** and **Y**, in order to characterize relations defined outside the window between the entities: 'The city of birth of **X** is **Y**'. If the system only takes into account the elements between the candidate pair, this kind of instances will not be extracted. However, including large contexts before the first entity and after the second one, may add noise and irrelevant elements to the classifier.

Second, the size of the intermediate window (between **X** and **Y**) also needs to be defined: If the window is small (5, 6, 7 tokens), the number of features and the complexity of the patterns will be reduced, but the classifier will have low coverage (since many instances will be excluded, as shown in Section 3.3). Increasing the window size will also increase the coverage of the system at the expense of growing the number of features. These windows will be empirically defined in Section 5.1.

4.3 Morphosyntactic

The feature space may be enriched by including morphosyntactic information, such as PoS-tags. PoS-tags can be added to the classifiers in two different ways: (i) by enriching the features based in lemmas (`Lem_PoS`) and by adding a new level of linguistic abstraction (`PoS`), representing a linguistic structure by means of morphosyntactic categories and not by lexical units. Thus, adding these features to

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

a classifier might help to improve the recognition of linguistic patterns. The PoS-tag features extracted from the example sentence are the following:

Sentence: **X** (born June 10, 1961) is an American **Y**
`Lem_PoS`: <FA>, <bear_VB>, <DATE>, <FC>, <be_VB>, <a_DI>, <american_ADJ>
`PoS`: <FA>, <VB>, <DATE>, <FC>, <VB>, <DI>, <ADJ>

4.4 Pseudo-syntactic

Syntactic analysis provides information about the function of different items in a linguistic structure. However, syntactic parsing may present some problems, since it is computationally expensive and usually its performance is still not very high, especially for languages other than English. Moreover, languages other than English often suffer lack of free tools for syntactic parsing.

Taking these assumptions into account, it is interesting to evaluate some features that, in some way, represent pseudo-syntactic information by means of encoding the position of the different elements in a sentence. Thus, the following features were evaluated during the experiments:

- `Bigrams`: bigrams of lemmas
- `Trigrams`: trigrams of lemmas

These features were encoded by using sequences of two (`Bigrams`) and three (`Trigrams`) adjacent lemmas:

Sentence: **X** (born June 10, 1961) is an American **Y**
`Bigrams`: <X_(>, <(_bear>, <bear_DATE>, <DATE_)>, <)_be>, <be_a>, <a_american>, <american_Y>
`Trigrams`: <X_(_bear>, <(_bear_DATE>, <bear_DATE_)>, <DATE_)_be>, <)_be_a>, <be_a_american>, <a_american_Y>

Other pseudo-syntactic features were also tested, such as intermediate and large lexico-syntactic patterns. However, they turned to be not helpful for any relation and language.

4.5 Syntactic

Finally, the last set of features are related to (pure) syntactic analysis. In recent years, many works have taken advantage of parsing in order to syntactically represent linguistic structures containing semantic relations.

Syntax encodes, in a deep linguistic level, the function of each element in a sentence. This way, the same syntactic functions could be extracted from sentences with very different surface structure. The following examples share the main syntactic structure, despite the fact that their surface form notoriously differ:

Example 1: **X** (born June 10, 1961) is an American $\mathbf{Y}_{(\text{singer})}$
Example 2: Many music journalists said that **X** was probably the best $\mathbf{Y}_{(\text{singer})}$ in America

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

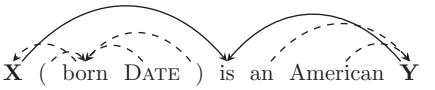


Fig. 4. Example of a shortest dependency path (with normal line) between **X** and **Y**.

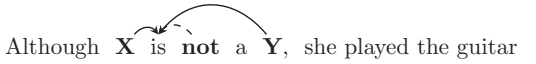


Fig. 5. Example of a shortest dependency path between **X** and **Y** that does not represent a semantic relation. In this example, crucial information (the negation of the verb) is lost.

Note that in the above sentences, **X** is the Subject of the verb *to be*, while **Y** (*singer*) is the attribute of the same verb.

In order to represent syntactic knowledge as features, three different ways were explored and evaluated:

- `Deps`: single syntactic dependencies
- `SDPs`: shortest dependency path (between **X** and **Y**)
- `Pths`: full dependency path

Single dependencies are triples that represent links between two related items (a head and a dependent) with a syntactic label. The first element of the dependency defines the syntactic function, while the head and the dependent are the second and third units, respectively.[7]

> *Sentence*: **X** (born June 10, 1961) is an American **Y**
> `Deps`: <Subj;be;**X**>, <Pt;bear;FA>, <Ad;**X**;bear>, <Obj;bear;DATE>, <Pt;bear;FC>, <Sp;**Y**;a>, <Ad;**Y**;american>, <Attrib;be;**Y**>

From this point of view, dependencies are similar to bigrams of lemmas, but built from (and providing) syntactic information. Note that in the previous example, the first and the last triples are crucial for defining the relation `hasProfession`, while the others add extra information. This observation inspired the next feature, called the shortest dependency path (SDP).

The SDP represents the shortest link between two entities in a dependency tree. Figure 4 shows (with normal line), the SDP between **X** and **Y** in the *Example 1* sentence. This way, the SDP of this sentence can be represented as the feature: <Subj;be;**X**//Attrib;be;**Y**>.

Although the SDP is often considered as one of the best ways for representing a relation between two entities in a sentence, some works have warned that important information (such as negation) might be lost by using SDP (Wu and Weld 2010). Figure 5 shows a sentence with the same SDP as the previous one, but not having the same semantic relation.

Therefore, a different path between the entities, using the full dependency path (`Pths`, which includes the dependents of the head elements in the `SDPs`), was evaluated. An example of this third syntactic feature can be seen below:

[7] Apart from the functions defined in page 12, Pt and Sp are shorter tags for Punctuation and Specifier.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Sentence: **X** (born June 10, 1961) is an American **Y**
Pths: <Subj;be;**X**//Pt;bear;FA//Ad;**X**;bear//Obj;bear;DATE//Pt;bear;FC//Sp;**Y**;a//Ad;**Y**;american//Attrib;be;**Y**>

This last feature will probably include irrelevant information in many cases, but it could also provide crucial elements ignored by the SDPs.

As in the previous categories, these syntactic features were evaluated alone and in several combinations. It is worth noting that the dependency parsers used to analyze the corpora (DepPattern, see Section 3.1) do not provide full coverage in some sentences. Since they are rule-based, some complex sentences were not totally parsed, so in these cases the SDP and/or the FDP were not extracted. However, these parsers performed better than Maltparser[8] trained with the Bosque[9] and AnCora[10] corpora for Portuguese and Spanish, respectively. Several evaluations show that the parsers used in this paper achieve about 88%/79%/83% (pt) and 85%/74%/79% (sp) of precision/recall/*f*-score (Gamallo and González 2013).

5 Experiments

This section shows the results of several experiments performed in order to know the effectiveness of each of the linguistic features previously defined.

First, some experiments concerning the size of the window between the entities are shown. Then, machine learning classifiers were trained using the features extracted from the linguistic space. The linguistic space is exploited using two strategies: first, by evaluating the performance of each type of feature individually, and second, in a bottom-up way, that is, by starting with the most basic features and, then, systematically including more complex features.

Experiments were carried out using the libsvm implementation of support vector machines (SVM) (Chang *et al.* 2011), since in preliminary tests it performed better than other algorithms such as Naive Bayes, Decision Trees, and Logistic Regression. In order to evaluate the ten relations (five for each language), a *one versus all* approach was applied.

For each classifier, 80% of the data were used for training and 20% for testing (all of them randomly selected). For setting the SVM parameters, ten-fold cross-validations were performed on the test sets.

In order to evaluate the performance of each classifier, the following measures were used: Precision is the number of correct positive decisions divided by the number of total positive decisions (true positives and false positives) made by each classifier. Recall refers to the number of correct positive decisions divided by the total number of positive examples in the test set. *F*-score is the harmonic mean of precision and recall.

Average results are presented using both micro-average and macro-average values. Micro-average is calculated by constructing a global contingency table with the

[8] http://www.maltparser.org
[9] http://www.linguateca.pt/floresta/corpus.html#bosque
[10] http://clic.ub.edu/corpus/ancora

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

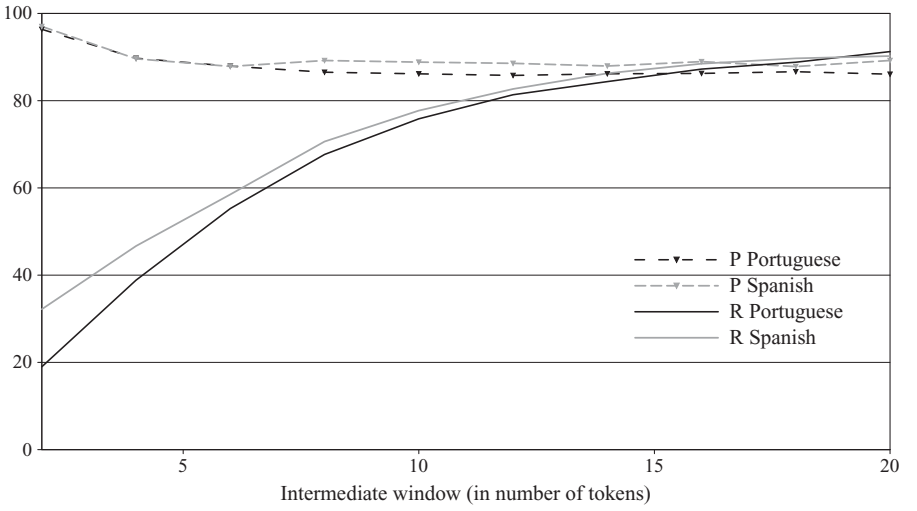


Fig. 6. Micro-average precision (P) and recall (R) of five classifiers for Portuguese and Spanish (one per each relation) *versus* intermediate window size (2–20, in number of tokens).

results for every relation, and then calculating precision, recall, and f-score using these sums. Macro-average values are the average of the individual results for each relation.

5.1 Windows sizes

Our first experiment consists in evaluating different windows sizes for the extraction task. Specifically, this experiment attempts to answer the following questions: What is the best intermediate window between **X** and **Y**? And what are the best windows for the sequences before the first and after the second entities?

This evaluation was performed with classifiers built with `Basic` and `Tok` features (Section 4) for all the relations and languages.

First, classifiers were trained using an intermediate window of two tokens, and incrementally adding two more tokens in every iteration until achieving a window of 20 elements. In these experiments, four tokens before the first entity and four after the second one were used.

Figure 6 contains the micro-average precision and recall of the five Portuguese classifiers as well as for the five Spanish systems. As this figure shows, using more than 16 tokens in the intermediate window does not improve the precision of the system. However, a large window allows the classifier to analyze more sentences, so the recall increases. Note that in Section 3.3, Figure 3 showed that most positive patterns occur with a small (< 8) intermediate window. For the rest of the experiments, an intermediate window of 16 was selected.

In the following, more tests were carried out in order to find the best size of the surrounding windows. The context of two entities **X** and **Y** consists of: the intermediate (between **X** and **Y**, already defined in 16 tokens) and the surrounding windows: before the first target entity (*previous*) and after the second one (*post*).

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Several experiments were performed using different sizes of these windows (from $0_{(previous)}/16_{(intermediate)}/0_{(post)}$ to 8/16/8).

However, the results of these tests show that changing the number of features in the *previous* and *post* contexts does not affect very much the performance of the system. In this respect, modifying from zero to eight the surrounding windows gave maximum differences of about 1% of *f*-score, and did not follow similar tendencies for each classifier.

Concerning the patterns, it is worth mentioning that the new structures obtained with larger surrounding windows are harder to be represented than those with shorter *previous* and *post* contexts. Therefore, the average results indicate that using windows of three tokens before and after the entities is enough for trading-off performance and efficiency (in number of features).

5.2 Feature effectiveness

In order to know the effectiveness of the different features the following strategy was applied: First, classifiers built with an individual type of feature (`Tok`, `Lem`, etc.) were used. The results of these experiments suggest what type of feature better represents the linguistic level they belong to (lexical, pseudo-syntactic, or syntactic). Then, several feature combinations were evaluated by training models with the best features of each level, trying to avoid information overlapping. Note that `Basic` features (described in Section 4.1) were incorporated in all the classifiers.

5.2.1 Semantic restriction and contextual subspaces

Before this main set of experiments, two baseline models were developed in order (i) to assess the importance of the semantic restriction of the candidate entities, and (ii) to determine the best way of representing the elements of the *previous* and *post* windows in the feature vector.

As Sections 3 and 4.1 noted, the approach followed in this work uses semantic knowledge in order to restrict the candidate pairs to be analyzed (classifying only PERSON–DATE pairs for `hasBirthPlace` and `hasDeathPlace`, and not for `hasBirthDate` or `hasProfession`, for instance). Therefore, two baseline models were used to evaluate how named entity recognition affects learning.

Baseline 1 (*B1*) classifiers use more candidate pairs than the other models, since they do not take advantage of semantic classification of proper nouns. Thus, negative pairs (which cannot belong to the testing relation: e.g., LOCATION–NUMBER for `hasBirthDate`) were automatically added to the training and testing sets. The number of candidates of each set is 5.2 (pt) and 4.9 (sp) times larger than the standard ones.

Baseline 2 (*B2*) systems differ from *B1* in the semantic restriction of the candidates, since *B2* applies semantic filtering on the candidate pairs.

Both *B1* and *B2* classifiers were built with `Tok` features extracted from the following windows: 3/16/3.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

The first two columns of Tables 3 and 4 show the results of the two baseline models for Portuguese and Spanish, respectively. Results are shown per relation as well as in micro- and macro-average values.[11]

As expected, the comparison of the *B1* to the *B2* results proves the effectiveness of the applied semantic filtering, which increases in more than 25% the micro-average *f*-score of the classifiers. Semantic restriction allows the models to not classify pairs that cannot belong to the target relation. Therefore, both precision and recall of all the classifiers and languages improve, except for the `hasDeathDate` relation, whose precision values decrease ($\approx$ 16%), but with much higher recall values ($\approx$ 60%).

B2 and *Lx1* models were trained with the same set of features, but the latter differs from the baselines on the way it represents the surrounding elements in the feature vector. Unlike baseline models, which mix all the features in the feature space, the following classifiers divide these features into three subspaces (*previous*, *intermediate*, and *post*). Therefore, the differences between *B2* and *Lx1* results (Tables 3 and 4) are due to the distinct representation of the surrounding elements in the feature vector.

Again, the performance of every classifier improves with the proposed feature representation. Contextual subspaces allow classifiers to better represent the sentences where the relation occurs, by identifying the relevance of the relative word order in a pattern.

It is worth noting that the differences between *B2* and *Lx1* results follow similar tendencies: in general, *f*-score results increase between 2% and 4% within the new scenario. However, those relations whose attribute (**Y**) is a DATE have a higher improvement (between 9% and 10% in Portuguese and 5% and 9% in Spanish). As it was shown in the pattern analysis (Section 3.3), these relations occur frequently in biographical structures containing both birth and death dates. So, the use of different subspaces helps these classifiers to disambiguate between many `hasBirthDate`/`hasDeathDate` candidate pairs. This strategy was applied in the training of the subsequent models.

5.2.2 Lexical and morphosyntactic level

The next experiment aims at knowing what is the best way for representing the lexical units in the sentences (Section 4.2). Three models were evaluated: *Lx1* (built with `Tok` features, as already mentioned), *Lx2* (built with `Lem`), and *Ms2* (based on `PoS` features), which contains morphosyntactic information.[12]

The results of these models (Tables 3 and 4) show, as hypothesized in Section 4.2, that the use of lemmas instead of tokens (*Lx2 versus Lx1*) generalizes the patterns, by reducing the data sparseness of the classifiers.

[11] BD refers to `hasBirthDate`; BP, `hasBirthPlace`; DD, `hasDeathDate`; DP, `hasDeathPlace`; and Pr means `hasProfession`. Additionally, *Ma* and *Mi* values are macro- and micro-average, respectively.

[12] *Lem_PoS* was also tested, but the results are omitted in this analysis because they are almost identical as those of *Lx2* models.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

This generalization involves little improvements in all the relations and languages (≈ 0.5%), except `hasDeathDate` in Spanish. Note, however, that the improvements against *Lx1* models were not statistically significant.[13] A deep analysis of these results showed that models with lemmas as *bag-of-words* correctly classify some candidates with less frequent verbal forms (e.g., 'nacidas'[14]).

Ms2 models are constructed with a further step of generalization, by substituting the lexical information of the sentences by their morphosyntactic representation.

Again, as the pattern analysis shows (Table 2), some relations heavily depend on biographical sentences, where punctuation takes an important role in the definition of the linguistic structures. These patterns are relatively well covered by `PoS` features, but removing the lexical information of the models involves a decrease in their general performance (≈ 3%), except for `hasProfession` relation, where representing adjectives (and other word classes less relevant for defining a relation) by their PoS-tag simplifies the feature space of the classifiers.

This set of tests confirms the validity of features based on lemmas (`Lem`) for representing the lexical units. The following experiments perform a similar analysis by evaluating the performance of models based on pseudo-syntactic features.

5.2.3 Pseudo-syntactic level

Two different models were developed using pseudo-syntactic information: *BiG*, built with bigrams of lemmas, and *TrG*, containing trigrams of lemmas.

Despite the results showing variation between relations and languages, in general bigrams of lemmas work better than models built using trigrams. In Portuguese, trigrams work better for `hasBirthPlace`, due to the great amount of patterns where the two target entities are covered by the same trigram feature <**X**_ (_**Y**_**Loc**>. Similarly, `hasProfession` results show less variation, since their patterns are usually well represented by both bigrams and trigrams of lemmas.

Best pseudo-syntactic models, however, do not clearly overcome *Lx2* classifiers, which perform slightly better in almost every scenario in Portuguese. Nevertheless, the *BiG* micro-average performance in Spanish is higher than *Lx2*, because *BiG* models only work noticeably better in `hasProfession` and `hasDeathDate`. In this case, only the improvements of *BiG* models for `hasProfession` were statistically significant.

5.2.4 Syntactic level

Finally, the effectiveness of syntactic parsing for the RE task was evaluated by comparing three different strategies: *Dep* models, which contain syntactic dependencies as the main feature; *Pth*, built with dependency paths; and classifiers using the shortest dependency path (*SDP*).

[13] Statistical significance was inferred at paired t-test, $p < 0.05$.
[14] Feminine plural participle of the verb 'nacer' (*to be born*) in Spanish.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Table 3. *Precision, Recall, and* F-*score per relation and model for Portuguese. Abbreviations are expanded in footnote* [11]

Rel		Model									
		B1	B2	Lx1	Lx2	Ms2	BiG	TrG	Dep	Pth	SDP
BD	P	69.3	83.8	93.4	93.8	88.8	91.3	81.3	70.3	59.2	60.7
	R	33.8	87.4	95.2	95.2	95.7	96.1	98.6	80.2	94.7	94.7
	F	45.5	85.6	94.3	94.5	92.1	93.7	89.1	74.9	72.9	74.0
BP	P	86.5	94.7	96.8	94.6	93.3	95.7	95.4	93.5	60.1	98.2
	R	72.3	91.4	94.1	95.7	92.2	95.7	96.9	96.1	96.5	63.3
	F	78.7	93.0	95.5	95.2	92.7	95.7	96.1	94.8	74.1	77.0
DD	P	92.3	73.9	82.5	87.0	88.5	82.3	85.0	58.8	0	100
	R	16.6	68.3	78.0	78.6	69.0	83.5	74.5	53.1	0	13.1
	F	28.1	71.0	80.1	82.6	77.5	82.9	79.4	55.8	0	23.2
DP	P	68.8	89.3	95.7	95.8	97.9	97.3	97.5	93.8	100	98.5
	R	16.6	85.5	87.6	88.5	71.0	86.1	81.6	81.6	4.2	19.9
	F	26.8	87.4	91.5	92.0	82.3	91.4	88.8	87.2	8.1	33.2
Pr	P	80.3	84.5	88.1	89.0	90.1	83.6	82.6	86.5	73.1	74.8
	R	76.8	92.1	92.8	92.3	91.9	96.1	96.5	91.7	82.0	83.3
	F	78.5	88.1	90.4	90.6	91.0	89.4	89.0	89.0	77.7	78.9
Ma	P	79.4	85.2	91.3	92.0	91.7	90.0	88.4	80.6	58.5	86.4
	R	43.2	85.0	89.5	90.1	83.9	91.5	89.6	80.5	55.7	54.9
	F	51.5	85.0	90.3	91.0	87.1	90.6	88.5	80.4	46.6	57.2
Mi	P	79.8	86.3	91.6	92.1	91.8	89.6	87.7	84.0	65.6	76.1
	R	49.0	87.2	90.6	91.0	85.2	92.3	91.0	84.4	59.9	59.0
	F	60.8	86.8	91.1	91.6	88.4	90.9	89.3	84.2	62.6	66.5

The three last columns in Tables 3 and 4 contain the results of the models that use syntactic information. Note that for some patterns, the dependency path and the SDP could not be extracted. This fact was caused for two main reasons: (i) lack of punctuation of some sentences, which involved parsing errors, and (ii) parser coverage, which cannot establish some dependencies in the entities path. Thus, 40% (pt) and 54% (sp) of the data do not have `Pths` features, while the `SDPs` were not extracted from 54% (pt) and 63% (sp) of the corpus.

Therefore, recall values of these classifiers are noticeably low (except for `hasProfession` and for `hasBirthDate` in Portuguese, whose patterns were better analyzed by the syntactic parsers).

In general, single dependencies do not have high precision values, unless they contain unambiguous knowledge (e.g., <Subj;nascer;**X**> and <em;nascer;**Y**>, where the verb 'nascer' means *to be born* in Portuguese), so the average results are worse than *B2* and *BiG* models.

Pth and *SDP* classifiers clearly depend on the coverage as well as on the patterns typology. The precision values of these models surpass *Lx2* and *BiG* in some relations (with values of 100% precision in several experiments). However, other

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Table 4. *Precision, Recall, and* F-*score per relation and model for Spanish. Abbreviations are expanded in footnote* [11]

Rel		Model									
		B1	B2	Lx1	Lx2	Ms2	BiG	TrG	Dep	Pth	SDP
BD	P	63.2	82.0	96.3	97.0	86.2	92.4	85.3	58.8	100	83.3
	R	25.8	89.8	93.4	95.2	86.2	94.0	97.0	70.1	1.8	12.0
	F	36.6	85.7	94.8	96.1	86.2	93.2	90.8	63.9	3.5	20.9
BP	P	74.5	90.0	90.8	94.6	88.4	91.8	85.9	86.1	43.1	100
	R	38.8	82.7	90.8	89.8	85.7	90.8	86.7	75.5	95.9	32.7
	F	51.0	86.2	90.8	92.2	87.1	91.3	86.3	80.4	59.5	49.2
DD	P	100	86.2	91.8	90.0	83.0	91.4	93.8	64.3	66.7	100
	R	9.4	79.0	84.1	83.7	77.6	84.6	77.1	55.6	0.9	10.4
	F	17.1	82.4	87.8	86.7	80.2	87.9	84.6	59.7	1.8	18.6
DP	P	0	93.1	95.7	98.0	97.4	97.5	96.0	90.4	100	92.9
	R	0	88.7	93.9	93.0	87.8	92.5	89.2	92.5	2.8	6.1
	F	0	90.9	94.8	95.4	92.4	94.9	92.5	91.4	5.5	11.5
Pr	P	87.9	90.4	93.1	93.5	93.9	91.5	89.7	87.4	75.3	84.5
	R	70.6	92.3	93.1	92.3	92.3	97.3	97.9	96.3	99.5	78.4
	F	78.3	91.3	93.1	92.9	93.1	94.3	93.6	91.6	85.7	81.3
Ma	P	65.1	88.3	93.5	94.6	89.8	92.9	90.1	77.4	77.0	92.1
	R	28.9	86.5	91.1	90.8	85.9	91.8	89.6	78.0	40.2	27.9
	F	36.6	87.3	92.3	92.6	87.8	92.3	89.5	77.4	31.2	36.3
Mi	P	84.6	89.0	93.5	94.2	91.3	92.5	90.2	80.5	68.8	85.9
	R	49.0	88.5	91.6	91.2	87.9	93.5	92.1	84.1	55.0	43.6
	F	62.1	88.7	92.6	92.6	89.6	93.0	91.2	82.3	61.1	57.9

patterns that contain ambiguous `Pths` and `SDPs` (or even that do not have these features) caused a lower performance.

Between these two models, *SDP* seems to have better precision than *Pth*, although both of them have recall problems, even in sentences with `Pths` and `SDPs` features successfully extracted.

So, this test indicates that using only features based on syntactic analysis (and even more depending on the parser accuracy and coverage) is not suitable for building RE systems. However, the high precision of some of the *Pth* and *SDP* models suggests that they could be good features to combine with other information based on lexical and/or pseudo-syntactic analysis.

5.3 Feature combinations

To optimize the combination of features in terms of performance and efficiency, different features from each linguistic level (lexical, pseudo-syntactic, and syntactic) were combined in order to find the best classifiers as well as to identify redundant features with information overlapping.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Table 5. *Precision, Recall, and* F-*score per relation, language, and model. Classifiers combine features from different levels of linguistic analysis:* L, *lexical (*`Lem`*);* B, *pseudo-syntactic (*`Bigrams`*);* P *and* S, *syntactic (*`Pths` *and* `SDPs`*, respectively)*

		Model					
		Portuguese			Spanish		
Relation		L/B	L/P	L/B/S	L/B	L/P	L/B/S
	P	92.2	93.4	91.4	97.0	97.6	97.0
`hasBirthDate`	R	97.1	95.7	97.1	95.8	95.2	95.8
	F	94.6	94.5	94.2	96.4	96.4	96.4
	P	96.1	95.0	95.8	95.7	95.8	94.6
`hasBirthPlace`	R	94.9	96.1	96.1	90.8	91.8	89.8
	F	95.5	95.5	95.9	93.2	93.8	92.2
	P	81.5	86.8	82.8	91.5	89.7	92.0
`hasDeathDate`	R	84.8	81.4	82.8	85.1	85.1	86.5
	F	83.1	84.0	82.8	88.1	87.3	89.2
	P	97.3	96.4	96.6	97.6	96.6	97.6
`hasDeathPlace`	R	87.0	88.5	86.7	93.9	93.0	93.4
	F	91.9	92.3	91.4	95.7	94.7	95.4
	P	88.4	89.7	90.8	94.5	93.1	94.4
`hasProfession`	R	93.6	92.1	93.4	94.8	93.1	95.3
	F	91.0	90.9	92.1	94.7	93.1	94.9
	P	91.1	92.3	91.5	95.2	94.5	95.1
`Macro-average`	R	91.5	90.8	91.2	92.1	91.6	92.2
	F	91.2	91.5	91.3	93.6	93.0	93.6
	P	91.5	92.5	92.2	94.9	93.9	94.9
`Micro-average`	R	91.0	91.4	91.8	92.9	91.9	93.2
	F	91.7	91.9	92.0	93.9	92.9	94.0

First, best lexical (`Lem`) and pseudo-syntactic (`Bigrams`) features were combined for building L/B models. Then, `Lem` were also used by training classifiers with `Pths` features (L/P).[15] Finally, three different levels of linguistic analysis were combined for building classifiers ($L/B/S$), by using the best possible combination of them: `Lem`, `Bigrams`, and `SDPs` features.

Table 5 contains the results of these three combinations for each relation and language, together with their macro- and micro-average values.

The first model (L/B), which combines features based on lexical and pseudo-syntactic analysis, performed statistically better than these features taken individually. Classifiers for Portuguese improve between 0.6% and 0.8% when compared to *Lx2* and *BiG*, respectively. In Spanish, the increase was of 1.27% and 0.9%. Furthermore, this combination involves improvements in both precision and recall of all the models (with the only exception of `hasDeathPlace` recall in Portuguese).

[15] `Pths` and `SDPs` features behave similarly in classifiers with just a single type of feature, so both of them were evaluated within `Lem` features. In these tests, the use of the full dependency parsing worked better than the `SDPs` ($\approx 0.3\%$).

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

So, lexical and pseudo-syntactic features, even built with the same base information (lemmas), are suitable combination for developing classifiers without computationally expensive resources such as syntactic parsers.

The next feature combination (L/P) was carried out by using lexical (`Lem`) and syntactic information (`Pths`). In Portuguese, the full dependency path also helps to surpass the performance of *Lx2* models in all the relations (with an increase of $\approx 0.4\%$ of micro-average f-score). These models also worked better (although just 0.2%) than the previous combinations (L/B).

In Spanish, the use of syntactic information together with features based on lemmas involves little improvements (0.2%) in the classifiers when compared to the use of lemmas as the unique type of feature. However, these combinations do not overcome models built with bigrams of lemmas (*BiG* and L/B).

Nonetheless, note that the Portuguese parser extracted more `Pths` and `SDPs` features than the Spanish one, as referred in Section 5.2.4.

Finally, the third combination ($L/B/S$) explores the use of `Tok`, `Bigrams`, and `SDPs` features in the same classifiers.

In Portuguese, this last combination works better than L/B ($\approx 0.3\%$) and L/P (0.07%) models, especially in those relations where *SDP* systems had achieved good results ($> 75\%$ f-score): `hasBirthPlace` and `hasProfession`. However, $L/B/S$ classifiers do not improve the performance of previous models in `hasDeathDate` and `hasDeathPlace`, in fact producing lower macro-average values than L/B and L/P systems.

In Spanish, the addition of syntactic features does not present noticeable improvements in the classifiers, although the average results are significantly better than those achieved by models built with just a single type of feature.

In sum, this last set of experiments demonstrates that combining different features that do not involve significant information overlap helps to improve the general performance of the classifiers. This means that complementary features permit better classifiers to be built. Taking into account that languages such as Portuguese and Spanish suffer lack of free or state-of-the-art tools for syntactic parsing, it could be interesting to evaluate these features with more mature systems.

5.4 Learning curve

Concerning the importance of the amount of training data used by the classifiers, Figure 7 shows the micro-average learning curve of the $L/B/S$ classifiers for Portuguese and Spanish.

These tests were carried out by using random sets of 25, 50, 75, 100, 200, 300, 400, 500, 600, and 637[16] training sentences. The curves indicate that the classifiers achieve their peaks with about 500 sentences. More training data gives little improvement in performance. However, other tests with relations that have more data (`hasBirthPlace` and `hasProfession`) suggest that the classifiers start

[16] Because 637 is the number of training sentences of the relation with the least data, `hasBirthDate` for Portuguese.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

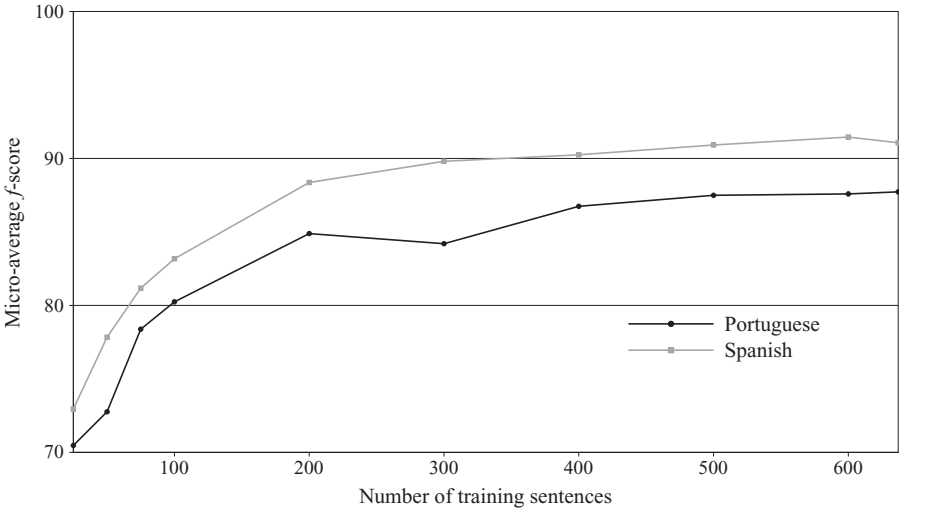


Fig. 7. Micro-average f-score of five classifiers for Portuguese and Spanish *versus* training data (25–637 pairs). Testing was performed with 200 random candidate pairs.

overfitting the training data with about 1,400 sentences, since their f-score results do not improve and sometimes decrease.

Finally, Figure 7 also shows that the learning curves have very similar tendencies for each language.

5.5 Error analysis

In order to know the main sources of errors made by the classifiers, a careful analysis of the false positives and false negatives of the $L/B/S$ results were carried out, including every misclassification of all relations and languages. The errors are grouped together by their typology. Moreover, the micro-average percentage (between *pt* and *sp* results) of each type of error was also calculated.

5.5.1 Ambiguity and parsing errors (≈43%)

This class of errors occurs when lexical and/or pseudo-syntactic features are not enough for representing the relation between two entities, and the SDP (which could disambiguate the pattern) is not extracted or not correctly built. The following Portuguese sentence (false positive for `hasBirthPlace`) exemplifies this kind of error:

Sentence: [. . .] filho de PER e de **X** nasceu no **Y_Loc** [. . .]
Trans.: [. . .] son of PER and of **X** was born in **Y_Loc** [. . .]
Pattern: [NP and of]$_{\text{previous}}$ **X** nascer em o **Y_Loc** [. . .]$_{\text{post}}$

5.5.2 Patterns without explicit information (≈25%)

Other misclassifications arise when the pattern (or even the whole sentence) does not contain explicit information concerning the target relation, thus producing a false

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

negative. The following example shows a pattern where the `hasBirthDate` relation (in Portuguese) between **X** and **Y_Dat** is not explicit, and the main lexical data is outside the scope of the extracted features:

Sent.: [. . .] teve sete filhos: PER em DAT, PER em DAT, **X** em **Y_Dat** [. . .]
Trans.: [. . .] had seven children: PER in DAT, PER in DAT, **X** in **Y_Dat** [. . .]
Pattern: [em DAT ,]$_{\text{previous}}$ **X** em **Y_Dat** [. . .]$_{\text{post}}$

5.5.3 *Less frequent patterns (≈15%)*

Some other positive pairs are misclassified (false negatives) when they appear in very infrequent patterns (with just one or two—or even zero—examples in the training corpus). Some of them might be correctly classified if they contain relevant features (such as a dependency path), but usually they are false negative instances:

Sent.: [. . .] **X**, que era apenas um homem da zona rural de **Y_Loc** [. . .]
Trans.: [. . .] **X**, who was just a man from the rural zone of **Y_Loc** [. . .]
Pattern: **X**, que ser apenas um homem de a zona rural de **Y_Loc**

5.5.4 *Useless shortest dependency path (≈14%)*

As shown in the previous section, SDPs are high-precision features. However, they can hurt the performance of a classifier when they are ambiguous for both true positive and true negative examples. The following pair of sentences are, respectively, a true negative and a true positive example of `hasProfession` in Spanish, represented by the same `SDPs`:

Sent. 1: **X** y un **Y_Occ** fueron dos de los homenajeados [. . .]
Trans.: **X** and a **Y_Occ** were two of the honorees [. . .]
Sent. 2: **X**, poeta y **Y_Occ**, recibió el patrocinio de [. . .]
Trans.: **X**, poet and **Y_Occ**, was sponsored by [. . .]
`SDPs`: <Coord;y;**X**//Coord;y;**Y_Occ**>

In this type of scenarios, the $L/B/S$ classifiers labeled as positive (for `hasProfession`) both of the sentences.

Besides that, `SDPs` extracted from biographical sentences—that often contain non-lexical dependencies between entities and surrounding punctuation—do not have high precision, since they appear indistinctly in positive and negative patterns.

5.5.5 *Other errors (≈3%)*

Finally, there are other less frequent sources of misclassifications, such as (i) lemmatization problems, and (ii) named entity recognition errors.

One of the few examples of errors produced by lemmatization includes the Portuguese participle 'morto' (*dead*), which was incorrectly lemmatized as 'matar' (*to kill*) instead of 'morrer' (*to die*). So, a `hasDeathPlace` pair such as the following was classified as negative for this relation:

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Sentence: **X**, morto em **Y_Loc** [. . .]
Trans.: **X**, dead in **Y_Loc** [. . .]
Pattern: **X** , matar em **Y_Loc**

Other less frequent sources of errors derive from the NER tool, which might incorrectly classify as a common noun a PERSON name, or vice-versa. In the following example, 'Belén' (LOCATION) was misrecognized as a common noun, so it was not a candidate entity for `hasDeathPlace`: '**X** (LOCATION, DATE - Belén, DATE).'

5.6 Discussion

This section briefly discusses the results of the performed experiments, taking into account the hypotheses referred to during the presentation of the features (Section 4) as well as some of the previous works analyzed in Section 2.

First, it is worth noting that filtering the candidate pairs by their semantic class allows the systems to increase their micro-average f-score by more than 25%, by improving both the recall and the precision. However, unlike other approaches (Kambhatla 2004; Zhou *et al.* 2005), the semantic classification of the other proper nouns in the patterns does not contribute to a better performance. Furthermore, the use of different subspaces for representing the surrounding elements of a pattern also improves the behavior of every classifier.

Then, it is also important to note that, once the candidate entities have been annotated with a NER, simple classifiers (those with only `Basic` and lexical features) are enough to achieve good precision and recall values (about 91% and 92% for Portuguese and Spanish, respectively). Thus, it can be inferred that (i) it is feasible to build high-quality proper noun ontologies with little manual effort, and that (ii) approaches based on supervised classifiers are useful for the BRE task.

Concerning the effectiveness of the features, first experiments show that generalizing the lexical units by means of lemmatization (`Lem`) improves the classifier performance. This fact agrees with tests carried out by Agichtein (2005), when he decides to use words (tokens) and stems. The use of PoS-tags, however, did not produce any improvement on the performed experiments (except for some patterns that depend little on lexical knowledge, such as '**Y_Occ X**,' as Mann (2002) noticed).

With regard to pseudo-syntactic features, the experiments performed in this paper showed that bigrams of lemmas were the best way to represent this level of analysis. Trigrams of lemmas also work well with patterns where the related entities are close together, but they still depend on other features for better representing the relation. Besides that, lexico-syntactic patterns had little impact on the classifiers.

Even though syntactic parsing might be computationally expensive and less accurate than other processing steps (lemmatization, NER), it provides useful information. Though syntactic features did not work well individually, their combination with lexical and pseudo-syntactic information allowed the systems to improve their classification. However, it is worth noting that some SDPs might be a source of errors if they represent entities linked by punctuation and/or coordination instead by lexical data.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Feature combination tests indicated that using complementary information obtained from different linguistic levels helps to develop better classifiers with less feature overlapping. In this respect, the best combinations for both Portuguese and Spanish were those based on lemmas, bigrams of lemmas, and SDPs. Apart from that, the performed experiments also show that using just lexical (`Lem`) and pseudo-syntactic data (`Bigrams`) permits high-precision classifiers to be built without using syntactic parsing.

In this regard, it is worth remembering that the parsers used in these experiments—but also other state-of-the-art parsers trained with free corpora—did not provide dependency paths for almost a half of the corpora, so the results should be considered as preliminary. Thus, the knowledge provided by syntactic features do not always add new information to the classifiers, since some dependency paths are similar to some bigrams of lemmas.

Therefore, some of the results of this work agree with Jijkoun *et al.* (2004) and Jiang and Zhai (2007), who had concluded that more complex features might hurt the performance of the classifiers (or increase the recall with a cost in precision). Nevertheless, syntactic information has also been reported as useful in previous works (Kambhatla 2004; Zhou *et al.* 2005; Mintz *et al.* 2009). Hence, it would be interesting to analyze the performance of syntactic features using different corpora and/or other parsers.

From the above, several conclusions can be made, which might be useful for improving the extraction of semantically related entities: First, a previous analysis of the patterns where the target(s) relation(s) occur can be useful for adapting the feature space of the classifiers. In this respect, works such as Garera and Yarowsky (2009) introduced several strategies designed *ad hoc* for biographical fact extraction.

Second, errors in lemmatization can be avoided by analyzing those lexical units that are crucial for a desired relation (e.g., 'to die' or 'to be born' verb for relations related to death and birth).

Third, if the patterns heavily depend on punctuation and/or coordination, syntax might not be very helpful, since parsers better represent syntactic relations established through lexical elements.

Finally, though the performance of the tested classifiers is good enough for accurately extracting instances of the evaluated relations, further work is needed in order to assess whether the methods presented in this paper are suitable for extractions with other relations and texts (Grishman 2010).

6 Conclusions and further work

Our work is an attempt to shed some light on the following questions concerning extraction of biographical semantic relations in Portuguese and Spanish: which are the more appropriate linguistic features to improve the performance of supervised machine learning approaches? Which are those allowing better combinations without feature overlapping? And what is the best way to obtain features for extracting different semantic relations from languages with less linguistic resources?

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

The experiments described in this paper can help us to outline some ideas to answer these questions. First, results showed that it is preferable to use lemmas instead of words, and that semantic knowledge is crucial for filtering the candidates of a classifier. Furthermore, combining lexical features with pseudo-syntactic information such as bigrams of lemmas improves the performance of the classifiers. Finally, syntactic parsing also provides useful knowledge to be combined with features from different levels of analysis, even if it does not involve great benefits in the systems.

Taking this into account, the results also pointed out that it is possible to build high-quality classifiers even for languages with fewer resources, since combinations of features obtained by lemmatization are often enough to train reliable RE classifiers.

Notice that this work dealt with several linguistic levels: lemmatization, PoS-tagging, named entity recognition, as well as pseudo-syntax and syntactic analysis.

Moreover, this paper showed that semi-structured knowledge available in Wikipedia is reliable enough to build a large amount of training data, without requiring bootstrapping techniques. This way, distant-supervision strategies can be adapted to rapidly obtain corpora for supervised classifiers, by reducing the time consumption of manually annotating corpora from scratch.

Semi-structured data such as infoboxes or category trees can be used as big repositories of *seed examples* to learn further semantic information on named entities, in the same way as lexical resources (e.g., WordNet) are often used as initial meaning repositories to deal with complex semantic phenomena related to lexical words: word sense disambiguation, anaphora resolution, etc.

Finally, let us mention the contribution of this work to the study of RE from languages other than English. Even if one of the main characteristic of the Web is its multilingual scope, most work on extraction of relations between named entities has been done only in the English version. This paper provides freely available annotated corpora with five biographical relations for both Portuguese and Spanish, as well as a careful analysis of the patterns where biographical relations occur. We hope there will be very soon more pieces of work on multilingual semantic extraction.

References

Agichtein, E. 2005. *Extracting Relations from Large Text Collections*, PhD Thesis. New York: Columbia University.

Agichtein, E., and Gravano, L. 2000. Snowball: extracting relations from large plain-text collections. In *Proceedings of the 5th Association for Computing Machinery Conference on Digital Libraries*, San Antonio, TX, USA, pp. 85–94.

Aguado de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., and Suárez-Figueroa, M. 2008. Natural language-based approach for helping in the reuse of ontology design patterns. In *Knowledge Engineering: Practice and Patterns*, pp. 32–47. Berlin: Springer-Verlag.

Akbik, A., and Broß, J. 2009. Wanderlust: extracting semantic relations from natural language text using dependency grammar patterns. In *Proceedings of the Workshop on Semantic Search (SemSearch 2009) at the 18th International World Wide Web Conference (WWW 2009)*, Madrid, Spain, pp. 6–15.

Banko, M., Cafarella, M. J., Soderl, S., Broadhead, M., and Etzioni, O. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, Hyderabad, India, pp. 2670–6.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Brin, S. 1998. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology (EDBT 1998)*, València, Spain, pp. 172–83.

Bruckschen, M., de Souza, J. G. C., Vieira, R., and Rigo, S. 2008. Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. In C. Mota and D. Santos (eds.), *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*, pp. 247–60. Linguateca.

Bunescu, R. C., and Mooney, R. J. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, pp. 724–31.

Bunescu, R. C., and Mooney, R. J. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp. 576–83.

Cardoso, N. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In C. Mota and D. Santos (eds.), *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*, pp. 195–211. Linguateca.

Chang, C., and Lin, C. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**(3): 1–27.

Chaves, M. S. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In C. Mota and D. Santos (eds.), *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*, pp. 231–45. Linguateca.

Costa, F., and Branco, A. 2012. Extracting temporal information from portuguese texts. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, pp. 99–105. Lecture Notes in Artificial Intelligence, vol. 7243. Berlin: Springer-Verlag.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. 2004. Web-scale information extraction in KnowItAll. In *Proceedings of the 13th International Conference on World Wide Web (WWW 2004)*, New York, USA, pp. 100–10.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Center, M. T. 2011. Open information extraction: the second generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, Barcelona, Catalonia, Spain.

Finkelstein-Landau, M., and Morin, E. 1999. Extracting semantic relationships between terms: supervised vs. unsupervised methods. In *Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure*, Dagstuhl Castle, Germany, pp. 71–80.

Fleischman, M., Hovy, E., and Echihabi, A. 2003. Offline strategies for online question answering: answering questions before they are asked. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, pp. 1–7.

Gamallo, P., and González, I. 2013. A compressing strategy for dependency parsing. Under review for *Revista Electrónica de Lingüística Aplicada*.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. 2012. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP (ROBUS-UNSUP 2012) at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France, pp. 10–18.

Gamallo, P., and González, I. 2011. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics* **16**(1): 45–71.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Garcia, M., and Gamallo, P. 2011a. An exploration of the linguistic knowledge for semantic relation extraction in Spanish. In P. Saint-Dizier and R. Mehta-Melkar (eds.), *Proceedings of the Joint Workshop FAM-LbR/KRAQ 2011. Learning by Reading and Its Applications in Intelligent Question-Answering at 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, Barcelona, Catalonia, Spain, pp. 7–12.

Garcia, M., and Gamallo, P. 2011b. Dependency-based text compression for semantic relation extraction. In P. Nakov, Z. Kozareva, K. Ganchev, and J. Hobbs (eds.), *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)*, Hissar, Bulgaria, pp. 21–8.

Garera, N., and Yarowsky, D. 2009. Structural, transitive and latent models for biographic fact extraction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athens, Greece, pp. 300–8.

Grishman, R. 2010. The impact of task and corpus on event extraction systems. In *Proceeding of 7th Language Resources and Evaluation Conference (LREC 2010)*, Valleta, Malta.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics* **2**: 539–45.

Hoffmann, R., Zhang, C., and Weld, D. S. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, pp. 286–95.

Jiang, J., and Zhai, C. 2007. A systematic exploration of the feature space for relation extraction. In *Proceedings of the Human Language Technologies/The Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2007)*, Rochester, NY, USA, pp. 113–20.

Jijkoun, V., De Rijke, M., and Mur, J. 2004. Information extraction for question answering: improving recall through syntactic patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 1284–90.

Kambhatla, N. 2004. Combining lexical, syntactic and semantic features with maximum entropy models for extracting relations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Catalonia, Spain.

Lin, D. 2003. Dependency-based evaluation of MINIPAR. *Treebanks: Building and Using Parsed Corpora* **20**: 317–29.

Liu, X., Nie, Z., Yu, N., and Wen, J. 2010. BioSnowball: automated population of Wikis. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, Washington, DC, USA, pp. 969–78.

Mann, G. S. 2002. Fine-grained proper noun ontologies for question answering. In *Proceedings of the 2002 Workshop on Building and Using Semantic Networks (SemaNet 2002)*, Taipei, Taiwan, pp. 1–7.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/AFNLP 2009)*, Singapore, pp. 1003–11.

Mota, C., and Santos, D. 2008. *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*. Linguateca.

Nagy, I., and Farkas, R. 2010. Person attribute extraction from the textual parts of web pages. In *CLEF (Notebook Papers/LABs/Workshops)*, Padua, Italy.

Nguyen, D. P. T., Matsuo, Y., and Ishizuka, M. 2007. Relation extraction from Wikipedia using subtree mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, Vancouver, Canada, vol. 2, pp. 1414–20.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Nguyen, T.-V. T., Moschitti, A., and Riccardi, G. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, vol. 3, pp. 1378–87.

Oliveira, H. G., and Gomes, P. 2010. Onto.PT: automatic construction of a lexical ontology for portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, Lisbon, Portugal, pp. 199–211.

Oliveira, H. G., Santos, D., Gomes, P., and Seco, N. 2008. PAPEL: a dictionary-based lexical ontology for Portuguese. In *Computational Processing of the Portuguese Language*, pp. 31–40. Berlin: Springer-Verlag.

Padró, Ll., Collado, M., Reese, S., Lloberes, M., and Castellón, I. 2010. FreeLing 2.1: five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, Valleta, Malta.

Pantel, P., and Pennacchiotti, M. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, NSW, Australia, pp. 113–20.

Pasca, M., Lin, D., Bigham, J., Lifchits, A., and Jain, A. 2006. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In *Proceedings of the National Conference on Artificial Intelligence*, Boston, MA, USA, vol. 21, pp. 1400–5.

Ravichandran, D., and Hovy, E. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, PA, USA, pp. 41–7.

Riedel, S., Yao, L., and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163. Berlin: Springer-Verlag.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Proceedings of the Atlantic Web Intelligence Conference (AWIC 2005)*, pp. 380–6. Lecture Notes in Computer Science, vol. 3528. Berling: Springer-Verlag.

Sánchez-Cuadrado, S., Lloréns, J., Morato, J., and Hurtado, J. A. 2003. Extracción automática de relaciones semánticas. In *2da Conferencia Iberoamericana en Sistemas, Cibernética e Informática (CISCI 2003)*, Orlando, Florida, pp. 41–7.

Sierra, G., Alarcón, R., Aguilar, C., and Bach, C. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology* **14**(1): 74–98.

Snow, R., Jurafsky, D., and Ng, A. Y. 2005. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems* **17**: 1297–304.

Soares, S., Martins, B., and Calado, P. 2011. Extracting biographical sentences from textual documents. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, Lisbon, Portugal, pp. 718–30.

Soler, V., and Alcina, A. 2008. Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español. *Terminology* **14**(1): 99–123.

Suchanek, F. M., Ifrim, G., and Weikum, G. 2006. LEILA: Learning to Extract Information by Linguistic Analysis. In *Second Workshop on Ontology Population (OLP2) at the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, NSW, Australia.

Sun, A., Grishman, R., Xu, W., and Min, B. 2011. New York University 2011 system for KBP slot filling. In *Proceedings of the Text Analytics Conference (TAC 2011)*, Gaithersburg, MD, USA.

Wan, X., Gao, J., Li, M., and Ding, B. 2005. Person resolution in person search results: WebHawk. In *Proceedings of the 14th Association for Computing Machinery International Conference on Information and Knowledge Management (CIKM 2005)*, Bremen, Germany, pp. 163–70.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press

Wu, F., and Weld, D. S. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, pp. 118–27.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., and Ishizuka, M. 2009. Unsupervised relation extraction by mining Wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/AFNLP 2009)*, Singapore, pp. 1021–9.

Zhao, S., and Grishman, R. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, MI, USA, pp. 419–26.

Zhang, M., Zhang, J., Su, J., and Zhou, G. 2006 A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, NSW, Australia, pp. 825–32.

Zhou, G., Su, J., Zhang, J., and Zhang, M. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, MI, USA, pp. 427–34.

Zhou, G., Zhang, M., Ji, D. H., and Zhu, Q. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, pp. 728–36.

https://doi.org/10.1017/S1351324913000314 Published online by Cambridge University Press