Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 279-292 recibido 05-12-2025 revisado 02-02-2026 aceptado 18-02-2026

# Wikipedia used as a semantic tagger: some preliminary results in Spanish

## *Wikipedia utilizada como etiquetador semántico: algunos resultados preliminares en castellano*

**Rogelio Nazar,**[1] **Irene Renau**[2]
[1]Pontificia Universidad Católica de Valparaíso
[2]Universitat Autònoma de Barcelona
Rogelio.Nazar@pucv.cl, Irene.Renau@uab.cat

**Abstract:** This paper describes a method based on data from Wikipedia for the automatic semantic tagging of common and proper nouns in context. We first predict the semantic category of each Wikipedia entry using a rule-based method that detects definition patterns, and then we generalize from there using a statistical model that associates semantic categories with elements of the entry. The evaluation of proper and common nouns in Spanish reveals a general precision of .82 and a recall of .77. One feature of the method is its conceptual simplicity and computational efficiency. The implementation is offered as open-source code and the data used in the study is in the public domain.
**Keywords:** ontology population, lexical taxonomy induction, word sense disambiguation, semantic tagging.

**Resumen:** El presente trabajo describe un método basado en los datos de Wikipedia para predecir la categoría semántica de nombres comunes o propios en contexto. Primero se intenta predecir la categoría de cada entrada de la Wikipedia mediante un sistema de reglas que detecta patrones definitorios, y luego se generaliza desde allí por medio de un modelo estadístico que asocia categorías semánticas con elementos de la entrada enciclopédica. La evaluación del etiquetado de nombres propios y comunes en castellano revela una precisión general de 82% y una cobertura de 77%. Una característica destacable del método es su simplicidad conceptual y eficiencia computacional. La implementación se ofrece como código abierto y los datos utilizados son de acceso público.
**Palabras clave:** poblamiento de ontologías, inducción de taxonomías léxicas, desambiguación automática, etiquetado semántico.

### *1 Introduction*

Semantic tagging is the task of assigning a semantic category from a given inventory to an expression, typically a single or multiword, proper or common noun, in a given text (Wilks and Stevenson, 1997; Asher and Pustejovsky, 2013; Bjerva, Plank, and Bos, 2016; Huang et al., 2022). The problem has a long standing tradition in natural language processing and is an essential part of named entity recognition and information extraction, with implications to other fields such as information retrieval and document categorization (Grishman and Sundheim, 1996; Tjong Kim Sang and De Meulder, 2003; Nadeau and Sekine, 2007; Kutbi, 2023).

Our particular interest in semantic tagging derives from a lexicographic project (Renau et al., 2019; Renau, Nazar, and Melanchthon, 2024) that follows the method of Corpus Pattern Analysis or CPA (Hanks, 2004; Hanks, 2013), which uses semantic types to organize the different patterns of use of a lexical unit. For instance, the fact that the subject or the object of a verb is a person or other type of entity can disambiguate between two patterns of such verb:

(1) *Mi madre bordó el pañuelo de mi boda.* ('My mother embroidered my wedding scarf')

(2) *La actriz bordó la interpretación.* ('The actress nailed the role')

ISSN 1135-5948 DOI 10.26342/2026-76-19 

Examples 1 and 2 are instances of the following patterns, respectively:

[[**Human**]] *bordar* [[**Artifact**]]

[[**Human**]] *bordar* [[**Performance**]]]

The elements in double square brackets are semantic types from the CPA Ontology, a bottom-up top-ontology, manually crafted by Hanks (2004), that contains ca. 250 very general concepts, such as [[Entity]], [[Property]], [[Event]], [[Food]], [[Drink]], etc., that can be used to semantically categorize most nouns in English, Spanish, and other languages. The two patterns just shown have the same transitive structure, and are different only by virtue of the semantic type in direct object position.

The technique is effective but the manual annotation of 250 – 1 000 concordances needed to analyze each lexical unit requires time, effort and specialized skills. Our motivation is therefore to automate at least parts of the process, and a critical step for speeding up the creation of CPA-based dictionaries is the semantic annotation of proper and common nouns of the corpus used for the analysis.

In previous research (Nazar, forthcoming), we described a different method for semantic tagging based on Wiktionary[1], which combined named entity recognition, ontology population and also word sense disambiguation, since the same word (e.g., *ratón* 'mouse') may require different semantic types according to the context of occurrence ([[Animal]] or [[Artifact]]). Results of that attempt left room for improvement because the dictionary covers mainly common nouns of the general vocabulary, with only limited treatment of proper nouns and specialized terminology.

In this context, the objective of the present paper is to improve semantic typing of words in corpus by expanding the variety of entities that can be recognized. This is achieved in this proposal with the use of a larger data resource (Wikipedia[2] instead of Wiktionary) and proposing a new method that turns this resource into a semantic tagger.

The proposal is divided in two parts. The first part consists in associating each Wikipedia entry with a semantic type in the CPA ontology. E.g., the entry for *Bertolt Brecht* corresponds to the category of [[Human]] and *Sitio de Sofía* ('Siege of Sofia') to [[Event]] (and not, say, [[Location]]). The second part consists of producing a statistical model to generalize the association of features in the Wikipedia entries with the semantic types such that it can also classify expressions that do not yet have an entry page.

Our main contribution is thus a new, fairly simple method, to produce semantic tagging in Spanish compatible with CPA methods. The simplicity of the proposed algorithm translates directly into computational efficiency and thus to its applicability to large corpora. It will also facilitate the replication of the method in other languages or its use with top-ontologies other than CPA's. Code, data and detailed documentation are publicly available at the project's website[3].

## *2 Related work*

### 2.1 Semantic tagging in the context of NLP

As already stated, semantic tagging is one of the oldest open problems in natural language processing, and it is also a fundamental task in many other operations. It is essential in any information extraction project, necessary for named entity recognition and categorization, and it implies word sense disambiguation tasks. But it is also related to other subfields such as hypernymy extraction, either from dictionaries (Chodorow, Byrd, and Heidorn, 1985; Guthrie et al., 1990) or from corpora (Hearst, 1992; Pearson, 1998; Meyer, 2001; Panchenko et al., 2016; Shwartz, Santus, and Schlechtweg, 2017; Sarkar, McCrae, and Buitelaar, 2018), as well as to other resource-building activities such as thesaurus generation (Grefenstette, 1994; Lin, 1998; Weeds and Weir, 2003; Bullinaria, 2008).

Semantic tagging is however more challenging than building resources because it not only implies the creation of some sort of lexicon and inventory of semantic tags, but also their application to a specific context. I.e., it is one thing to create a database associating words with semantic tags, but another very different one is to decide, for the specific oc-

[1] https://www.wiktionary.org

[2] https://www.wikipedia.org

[3] The materials are currently hosted on the website of Project Text·a·Gram, which offers different tools for text analysis (<http://www.tecling.com/textagram>). The prototype described in this paper, which, for the time being, is called *Wicacho*, also has its own web-demo: <http://www.tecling.com/wicacho>

currence of a word, what is the appropriate tag for the occasion. This is of course an instance of the general problem of word sense disambiguation, another of the traditional problems in computational linguistics (Lesk, 1986; Yarowsky, 1992; Gale, Church, and Yarowsky, 1992; Navigli, 2009; Camacho-Collados and Pilehvar, 2018; Wiedemann et al., 2019).

Historically, semantic tagging consolidated by the 1990s in venues such as the Message Understanding Conferences (MUC), Computational Natural Language Learning (CoNLL) and Automatic Content Extraction (ACE) (Grishman and Sundheim, 1996; Li et al., 2023). Competitions were organized in which participants had to propose methods to process input text and 1) identify single or multiword expressions; 2) assign them a semantic tag from some inventory.

In the first versions, the extracted entities were limited to proper nouns and the inventories were rather simple, with limited possibilities such as persons, locations and organizations. However, the task gradually became more and more complex, and the inventories grew to classify all types of expressions in a wide range of specific semantic categories, such as dates, numbers, quantities, properties, and so on (Sekine and Nobata, 2004; Nadeau and Sekine, 2007; Nadeau, 2007; Ling and Weld, 2012; Abzianidze and Bos, 2017; Li et al., 2023). In most recent proposals, coarse-grained categories such as [[Human]] have been replaced with fine or ultra-fine types, such as [[Banker]], [[Politician]], [[Lawyer]], etc. (Choi et al., 2018; Xiong et al., 2019). This fine-grained semantic detail would be unnecessary for our purposes in CPA because it would not help disambiguate patterns.

## 2.2 Semantic typing in the context of CPA

As earlier mentioned, the CPA ontology has ca. 250 semantic types which are used to differentiate the patterns of use and nuances of meaning of lexical units. According to Hanks (2013), these patterns are conventional linguistic norms established by language use. Norms are different from the most creative or innovative linguistic expressions, called 'exploitations' in Hanks' theory. Examples 1 and 2, in the introduction, are normal uses of the verb *bordar* in Spanish. In contrast, example 3 would be a metaphorical exploitation.

(3) *La primavera borda el campo con espárragos.* ('Spring embroiders the countryside with asparagus')

CPA is the application of this theory of norms and exploitations to practical lexicographic work, and semantic types are an essential component of this technique. Typing the arguments of predicates in a sufficiently large corpus makes different pattern begin to emerge.

Despite being a promising lexicographic framework, many CPA practitioners working in different languages have however acknowledged that the manual annotation of corpus required for the analysis is an obstacle for the efficient development of dictionaries, and this has been a motivation to automate at least part of the process (Pustejovsky, Hanks, and Rumshisky, 2004; Ježek et al., 2014; Colman and Tiberius, 2018; Marini and Ježek, 2019; Giacomini and DiMuccio-Failla, 2019; Renau, Nazar, and Melanchthon, 2024).

## 2.3 Available solutions for semantic tagging

There are currently many available solutions for semantic tagging in Spanish, each offering different advantages and disadvantages. Some commercial applications, such as Babelfy[4] (Moro, Cecconi, and Navigli, 2014); Dandelion[5]; displaCy[6] or TextRazor[7] seem effective, but non-viable for their application to large corpora. The open-source community has also offered some alternatives. Interesting proposals are DBpedia[8] (Daiber et al., 2013); the NER tagger from the FlairNLP[9] framework (Akbik, Bergmann, and Vollgraf, 2019); the NLTK[10] library (Bird, Loper, and Klein, 2009); Polyglot[11] (Al-Rfou et al., 2014); PyMusas[12] (Rayson et al., 2004; Piao et al., 2016) and StanfordCoreNLP[13] (Finkel, Grenager, and Manning, 2005).

An extensive and detailed evaluation of each of these alternatives is beyond the scope

[4] http://babelfy.org
[5] https://dandelion.eu/semantic-text
[6] https://explosion.ai/demos/displacy-ent
[7] http://www.textrazor.com
[8] http://spotlight.dbpedia.org
[9] https://flairnlp.github.io
[10] https://www.nltk.org
[11] https://polyglot.readthedocs.io
[12] https://ucrel.lancs.ac.uk/usas
[13] https://stanfordnlp.github.io/CoreNLP

of this paper. However, a general examination reveals some limitations. FlairNLP, for instance, shows robust performance but, as a NER tagger, it has a reduced tagset (mainly *person*, *place* and *organization*). Others, like Babelfy and PyMusas, have a larger tagset but suffer from WSD problems, as well as low recall given that they cannot deal with elements not already listed in their knowledge base. DBpedia is interesting because of its use of Wikipedia, but to a certain degree it has the same problems as the others plus some complexity for integration and maintenance, with many different modules, libraries and large disk space requirements.

Another problem with the currently available methods for semantic tagging in Spanish is the lack of interoperability, as each system provides a different data format. We leave for future work the possibility of trying to adapt or standardize their output. Given the available options, we judged that the development of a new method, specific for the CPA Ontology tagset, was justified if simple enough for the application to different languages and with minimal computational cost.

The most recent tendencies are far from the idea of simple solutions. There are reports of successful application of deep learning methods to semantic tagging (Li et al., 2023; Hu, Hou, and Liu, 2024), but these systems are characterized by their lack of explainability, the great complexity, and their energy consumption, particularly the transformer-based large language models (LLMs). Our own attempts with them, in addition, have so far been unsuccessful. Results were inconsistent or incorrect as these systems tend to create their own semantic tags instead of using the ones mentioned in the prompt. Even if successful, and leaving aside questions of scientific explainability or interpretability as well as their non-deterministic nature, the cost of scalability to annotate large corpora would still remain insurmountable.

## *3 Methodology*

As already stated in the introduction, we propose a methodology for semantic tagging based on data from Wikipedia, which improves upon an earlier attempt based on Wiktionary. In the earlier attempt, we were able to manage the tagging of common nouns relatively well, but the coverage of proper nouns and specialized terminology was insufficient, mainly due to the fact that a lexical database derived from a conventional dictionary is not large enough to include the great diversity of entities that may appear in a text. The need to reduce the ratio of out-of-vocabulary units lead us to try a new method this time using Wikipedia instead of Wiktionary.

In this section we explain thus the adaptation of the encyclopedic material (3.1), the alignment of each entry to our tagset (3.2), the generalization of this alignment via statistical modeling (3.3) and the application of these resources to the tagging of a particular text (3.4).

### 3.1 Preparation of materials

In order to use this encyclopedia effectively for our purposes, we downloaded a recent version of the resource from the Spanish Wikipedia Dumps site[14]. For efficiency, we used only the index file, a reduced version of the resource which includes only the title and the first paragraph of each page, which is the one usually containing the basic information about the entity.

The first part of the process is to convert the original XML code into a CSV table. Each entry of the Wikipedia index file has three fields: the title, the URL, and the abstract plus, in some cases, additional links. Figure 1 shows an example of an entry, in this case relating to the Covarona cave, located in Northern Spain.

```
<doc>
<title>Wikipedia: La Covarona</title>
<url>https://es.wikipedia.org/wiki/La_Covarona</url>
<abstract>La Covarona es una cueva prehistórica situa
barrio de Llueva, en San Miguelde Aras, del municipio
en Cantabria (España). Su nombre procede del euskera
"Coba" (cueva) y "ona" (buena).
</abstract>
<links><sublink linktype="nav">
<anchor>Véase también</anchor>
<link>https://es.wikipedia.org/wiki/La_Covarona#Véase
</link></sublink>
</links>
</doc>
```

Figure 1: Example of an entry in the Wikipedia XML Dump file.

We extracted only the content of the XML tags 'title' and 'abstract', which results in a matrix $M_{x,y}$ with 1,919,985 rows, with an index $x$ for the title and $y$ for the available text, which not always looks as polished as in this example[15].

[14] https://dumps.wikimedia.org/eswiki/

[15] In 24% of the entries there is only incomplete text

### 3.2 Alignment of Wikipedia to the CPA Ontology

At this point of the process we assigned each Wikipedia entry a unique semantic type from the CPA Ontology. This tag set was originally developed in English, but we manually translated it to Spanish for the purpose of this study[16]. Arguably, this would be a form of ontology population, and it consists in the classification of the almost two million entities in matrix $M$ in the different available categories of the CPA Ontology. The idea is that the process results in a tree-like structure $C$, i.e. a directed acyclic graph where each leaf node has a unique ascending path to the vertex.

For a core set of basic ontological categories ([[Human]], [[Location]], [[Eventuality]], [[Property]], etc.), our strategy was to set up a battery of tests. If all tests fail, then we apply a set of pattern identification techniques to identify definitions, similar to those used by Hearst (1992) to obtain hypernymy links from corpora.

In the case of the test for [[Human]], there are rules to detect features that are typical of human subjects. These rules apply to the name of the subject as well as to the available paragraph in the Wikipedia entry. In the case of the name, bearing a common first name and/or surname is a strong indication of human nature. For this we compiled a list of both types of names from the same resource. This is done iteratively from a few common seed names. Say, for instance, for the first name *Lorena*, we obtain a ranked list $L$ of elements appearing immediately after this name, sorted in order of decreasing frequency: $L = \{l_1, l_2, l_3...l_n\}$. From this set, we would define a subset $S$ of the $k$ most frequent elements (1), such that $\{Sánchez, Méndez, Pérez...\} \subset S$.

$$S = \{l_{(i)} \mid 1 \leq i \leq k\} \quad (1)$$

For each element of $S$ (e.g., *Sánchez*), the same process is repeated this time to obtain first names from matrix $M$ (i.e, now elements appearing before the searched element), iterating the process[17].

Other indications that the name designates a person can be found in the accompanying paragraph $y$. For simplicity, rules that require deep syntactic analysis are left for future research. Instead, we apply pattern matching of sequences such as those shown in table 1, which try to capture the typical context where a person name is used.

A similar strategy was followed for other basic category, like [[Location]]. The same procedure was followed to obtain a list of place indicators from the same matrix $M$. From a seed list of location names, we obtained elements that frequently co-occur with those names, composing in this way a list of location triggers (e.g. *Bahía, Ciudad, Torre, Península* –Bay, City, Tower, Peninsula). In this way, if $y$ contains for instance "la ciudad de $x$..." ('the city of'), $x$ would be tagged as [[Location]]. Similarly, for the semantic type [[Event]], the same strategy is followed, with a different list of manually created patterns.

When these tests failed, a battery of patterns was applied to $y$ to extract a hypernym $h$ for entity $x$. For illustration, Table 2 offers some examples of these rules, typed as regular expressions. The dollar sign indicates keywords, like cardinal and ordinal numbers and forms of the verb *ser* ('to be') in Spanish. The 'various' keyword refers to a set of words that designate groups (*clase, conjunto, especie, familia, forma, género, grupo, orden, pieza, serie, subclase, subespecie, tipo, variedad* – 'class, set, species, family, form, genus, group, order, piece, series, subclass, subspecies, type, variety'.)

If $h \in C$, then a tag for $x$ has been found. Otherwise, a new hypernym for $h$ is obtained, recursively, to test if the new, more general element, is now contained in tagset $C$. This recursion is limited to three cycles in order to save time.

In the case of multiword expressions, a last resource when a tag cannot be found is to retry using the head of the phrase. Without parsing, in Spanish one can expect the first element of the sequence to be the head. If yet no tag is found, the element receives the temporary tag of $UNKNOWN$, and it may be tagged at a later stage in the process, as explained in the following subsection.

(just a few characters) or simply no text at all.

[16]The translation is also available at the website of Project Kind: http://www.tecling.com/kind

[17]Manual revision of the lists between iterations would be impractical due to their large size, so a certain degree of error must be assumed. We obtained ca. 220 000 first names, 180 000 surnames, 13 000 locations and 20 000 names of organizations.

| **Category** | **Triggers** |
|---|---|
| Human | *abogad, académic, activista, actor, actriz, ajedrecista, alfarer, almirante, angel, animador, antipapa, antropólog, árbitro, arqueólog, arquitect, artista*, etc. |
| Location | *Superficie, Provincia, Comunidad, Municipio, República, Puerto, Castillo, Palacio, Basílica, Parque, Bahía*, etc. |
| Organization | *academia, agencia, agrupación, alianza, asociación, cadena, canal, comité, cooperativa, corporación, emisora, escuela*, etc. |
| Cultural product | *novela, película, canción, relato, (corto, largo)metraje*, etc. |
| **...** | **...** |

Table 1: Examples of patterns used to test general categories.

| |
|---|
| nombre (dado\|con el que (se hace referencia\|se conoce\|se suele conocer\|denomina\|designa\|designaban)) al? (l[ao]s?\|una?)? ((antigu\|ampli\|extens)[ao]s\|$ord\|$card)? ? |
| .*(se conoce como\|se define\|se denomina\|se entiende por\|se le llama\|se llama) .+? al? (una?\|l[ao]s)? |
| ^($be) +(el\|las?\|una?) (pequeñ[ao]\|$ord) ([^$word]+) |
| ^un[oa] de l[ao]s (poc[as]s\|primer[ao]s\|principales\|tipos de\|($card)) ([^$word]+) |
| es cualquier tipo de\|es aquella) (de\|que\|en)? ?([^$word]+) |
| nombre (común )?de ($various )? |
| ... |

Table 2: Examples of patterns used to extract information from Wikipedia abstracts.

### 3.3 Development of a statistical model for semantic tagging

The rules described in the previous section are precise in general, but they also tend to present relatively low recall, that is, many of the elements end up with the category $UNKNOWN$. This occurs either because no pattern was found or because there was no match between the hypernym obtained from the abstract and tagset $C$.

This is why we need a subsequent part of the process, where we derive a statistical model from the results obtained in the previous part. Using the result of the previous process as input, we created a two column matrix $L$ to associate semantic categories with three types of data: 1) words (single word types of character length $> 3$) frequently found in the entry names; 2) words frequently found in the abstracts and 3) derivational suffixes (sequences of 3 – 6 characters at the end of the word) frequently found in the head of the entry names, since these segments are associated with semantic categories (Light, 1996; Namer, 2002). For instance, in Spanish the suffix *-miento* denotes processes or events, while *-dad* is associated with properties, etc.

An input entity $x$ will thus receive the semantic tag that offers the greatest score in equation 2, where $T(x)$ represents $x$ with its features.

$$C[x] = \max_{i=1}^{|L|} \left( \prod_{j=1}^{|L_i|} 1 + |T(x) \cap L_{i,j}| \right) \quad (2)$$

Even when the statistical method is used to extend and generalize over the one based on rules, both can also be used in parallel to reinforce confidence when they produce the same result. Having both running in parallel may also be useful to obtain more information. For instance, in the case of the input name such as *La Covarona*, the semantic type resulting from one method is [[Location]], while the other produces [[Cave]], i.e., a landscape feature that is a more specific type of place.

### 3.4 Semantic tagging of a text

Once all the necessary materials have been created during the preparation steps described in the previous subsections, it is now possible to produce the semantic tagging of the texts of a corpus. For any arbitrary input text, the two main tasks to undertake are first to identify the entities in the text and then to assign the entities a semantic type.

#### 3.4.1 Identification of the entities

In the current state of our project, the phase of entity spotting (i.e., the detection and proper delimitation of the entities mentioned

in the texts) is the first step before semantic tagging. Thus, any given input text is scanned from top to bottom to detect entities, and the identification is carried out with procedures informed by grammatical data. The main difficulty is the treatment of multiword expressions, because one needs to detect the beginning and the end of each sequence.

In an effort to keep the methodological proposal as simple as possible, no attempt was made to use full syntactic parsing of the sentences in the analyzed corpus. Instead, we applied only a POS-tagger that for each token produces a lemma and a POS-tag. For this we used UDPipe[18] (Straka and Straková, 2018).

Without full syntactic parsing, one can only work with the sequences of POS-tags, performing a form of shallow parsing or chunking aimed at the detection of sequences that conform a particular type of noun phrase. The system admits as valid units only nouns, proper nouns, adjectives, definite articles and the preposition *de* ('of').

A sequence begins when a noun is found, and it ends when a non valid element is found, like a punctuation sign or a conjunction, etc. Eventual prepositions or articles remaining in the last position of the sequence are discarded. Once an entity is closed and saved, the process resumes from the same point onwards until the end of the text is reached.

The above method will produce valid sequences such as *República Federal de Alemania* ('German Federal Republic') or *dióxido de carbono* ('carbon dioxide'). Of course, without grammatical information, the method is error prone as it is blind to the boundaries between arguments and predicates and between different arguments of a predicate.

#### 3.4.2 Semantic typing of the detected entities

Once with a list of entities extracted from the text, the semantic typing procedure starts analyzing one unit a time. Entities are subject to a battery of tests very similar to those explained earlier (section 3.2). The final decision for a semantic type is the result of a combination of factors, computed by a voting scheme.

Factor 1 consists of the selection from the available pages. As described earlier, the data from Wikipedia is here organized in the form of a table $C$ in which every unit $x \in E$ is assigned a semantic tag $C[x]$. This is straightforward when there is only one page in Wikipedia for such name ( $|M_x| = 1$). E.g., $\exists x = \mathit{Covarona} : C[x] = \mathit{Location}$. This, of course, is often not the case, and in those occasions, a WSD routine is invoked.

Drawing inspiration from early methods in the field (Lesk, 1986), the disambiguation of a unit $x$ proceeds by calculating the vocabulary match between the text $T(x)$ where it appears and the paragraphs corresponding to the different pages or 'senses' of entity $x$ in Wikipedia. For simplicity, the vocabulary intersection is computed considering only single content words. According to equation 3, a first factor $F_1(x)$ will select from a set of available types ($D(x)$) the one that shows the greatest vocabulary overlap.

$$F_1(x) = \arg \max_{s_i \in D(x)} |T(x) \cap M_{x,i}| \quad (3)$$

Factor 2 operates in the same manner as described in subsection 3.2. Here, a semantic type gains points if triggers, described before as a keyword associated with a semantic type, are found in the name of the entity or in the immediate vicinity of the contexts of occurrence of the entity, using a symmetric context window of three tokens. The procedure is the same as defined in equation 3, only that now the comparison is between the target text $T(x)$ and each of the set of triggers associated with each semantic type.

Factors 3, 4 and 5 are based on the statistical model described in subsection 3.3, i.e., the model $L$ that associates semantic types with vocabulary items found in the different Wikipedia entries corresponding to those types. Factor 3 measures the match at the entry level, i.e., words in the name of the entity that may reveal the category. Factor 4 is similar but the match is in words occurring in the immediate vicinity of the entity's mentions. Factor 5 measures the match between the suffixes of the noun (or the head of the noun phrase referring to a given entity) and the suffixes associated with a semantic type.

They are applied as expressed in equation

[18] This software also offers full parsing, but it takes longer, produces more data, and the required post-processing of the resulting syntactic data has its intricacies. Full parsing also further complicates future possibilities of replicating experiments in other languages, especially those with less resources.

3, the difference being that this time vocabulary match is not computed comparing $T(x)$ and the definitions in Wikipedia ($M_x$) but, instead, the vocabulary in $T(x)$ is compared with the vocabulary associated with each semantic type in model $L$.

The final decision is made by simple voting (equation 4), in which a rank $R(x)$ for entity $x$ presents the most frequently selected semantic type $s$ within the factors earlier mentioned.

$$R(x) = \arg\max_{s \in C} f(s) \tag{4}$$

If there is no semantic type selected or if the best one has the same points as the second best, then no type is selected and the result is the tag *UNKNOWN*. Optionally, a 'conservative' setting can be configured, and in this case there will only be a response if a given semantic type is selected by at least $k$ factors.

## 4 Results

For the evaluation of the method, we conducted a study tagging a random sample of 200 Wikipedia pages in Spanish. The sample was constrained to be about historical figures (military personnel from the $19^{th}$ Century), with frequent cases of ambiguity, e.g., places that receive their name from a military figure. The corpus has a total size of 395,284 tokens.

From the tagging result, we draw a random sample of 1000 entities and evaluated precision, recall and F1 in the classification. Two annotators carried out the evaluation separately, leaving 500 trials for the calculation of intercoder agreement. There were a total of 54 disparities between the two, making 89% agreement. Cohen's Kappa resulted in a value of 0.72, which falls in the category of substantial agreement (0.61–0.80).

The evaluation is divided in two steps. The first one is to analyze the result of the chunking process explained in subsection 3.4.1. A total of 57 out of 1000 units were found to be incorrectly segmented. In most cases, the reasons that explain these failures are those already discussed in the methodology, i.e., the lack of full syntactic parsing. Example 4 shows the case of the incorrect sequence *empresa la prisión del cuadro* (Lit., 'company the prison of the frame'), with the typical confusion between arguments of the predicate, *empresa* ('actions') and *prisión* ('imprisonment').

(4) *...ha sido igualmente fruto de nuestra* ***empresa la prisión del cuadro*** *de oficiales que formaban la escolta...* ('...the imprisonment of the officers who formed the escort has also been the result of our actions...').

Cases like these are excluded from the reminder of the evaluation, which consist of rating the performance of the semantic typing process. Table 3 presents a summary of the evaluation figures. There we show the results of two configurations with more or less confidence in the results, according to the parameter explained in subsection 3.4.2.

| Measure | Normal | Conservative |
|---|---|---|
| TP | 630 | 442 |
| FP | 141 | 18 |
| FN | 189 | 481 |
| Precision | .82 | **.96** |
| Recall | **.77** | .48 |
| F1 | **.79** | .64 |

Table 3: Evaluation figures in a random sample of 1000 trials.

In this table, a true positive (TP) means a case where an entity receives a correct tag; a false positive (FP) is a case where an entity receives an incorrect tag and a false negative (FN) is a case where an entity receives no tag. Precision (5) is the proportion of cases in which the algorithm returned a correct category over the times in which it produced a result. Recall (6) is in turn defined as the proportion of correct results over the sum of true positives and false negatives. F1 (7) is the harmonic mean between precision and recall.

$$pre = \frac{TP}{TP + FP} \tag{5}$$

$$rec = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{7}$$

We consider that, at these early stages of the project, these performance figures are promising, and a considerable improvement over our previous attempt. The 'conservative' configuration suffers considerable loss of

recall but the increment in precision is interesting, as this result could be used to collect training examples for other statistical algorithms.

A more detailed analysis of results reveals systematic causes in many cases and suggests ideas for future versions. Table 4 shows a few cases of entities, with their classification by the algorithm and our evaluation.

Regarding false negatives, the most frequent case was those of single word lexical units of the general vocabulary, which tend to be less represented in Wikipedia (e.g. *adolescente, bulto, franqueza*, etc. ('adolescent, package, frankness'). It is worth noting however that this is precisely the type of nouns we could process relatively well with the previous method.

Regarding false positives, a frequent cause of errors was problems with the POS-tagger, which tends to confuse common nouns with proper nouns when they appear with initial upper case letter, as they normally do at the beginning of the sentence. When this occurs, the lemma assigned is incorrect. This affects especially the case of common nouns when they appear in plural (e.g. *Diccionarios, Fallecidos, Generales*, etc. ('Dictionaries, Deceased, Generals').

Another frequent cause of errors was failure to disambiguate. It would be the case of *explorador* ('explorer'), incorrectly tagged as [[Vehículo]] ('vehicle') instead of [[Persona]] ('person'). This circumstance has been further aggravated in this case because of the particular characteristics of the evaluation corpus, in which the persons named in these texts, being historical figures, often end up giving their names to places, a common transit from anthroponym to toponym. This explains the incorrect assignment of the category of 'Place' to *Amán Rawson* in table 4, possibly confused with the city of Rawson, in Chubut, Argentina. This ambiguity is systemic with surnames such as *Dorrego, Balcarce* or *San Martín*, which now are the names of cities, districts, parks and avenues. In the same vein, the corpus presents many direct quotations of text from those times, and presents instances of military jargon and also archaic formulations. This would be the case of *asiento*, in table 4, incorrectly classified as [[Mueble]] ('furniture').

The statistical modeling described in section 3.3 was effective in general but also contributed with some errors. For instance, it correctly generalizes over cases like *Departamento de Potosí* or *Departamento de la Paz*, which designate places, and classifies successfully *Departamento de Rivadavia*, but it also over-generalizes, to incorrectly treat *departamento de medicina* ('medicine department') as a place.

In other cases, this module leads to inadequacies, such as *círculo de amigos* ('circle of friends) tagged as [[Organización]] ('organization') when it should be [[Human Group]]. A similar case is *Corona de España* tagged as [[Lugar]] ('place') instead of [[Organización]], in a segment like *había ya obtenido su independencia de la Corona de España* ('had already gained independence from the Spanish crown').

Cases of phraseology were also among the most frequent error typologies. This is the case of *al aire libre* ('in the open') and many other locutions or fixed expressions in Spanish such as *a pesar de* ('despite'), or *poner freno a* ('to clamp down on'), which explains the incorrect assignment of the semantic type [[Device]] to *freno* in a fragment like *tenía como objetivo poner un freno a la expansión del unitarismo* ('aimed to put the brakes on the expansion of Unitarianism'). This is a reminder of the importance of having linguistic resources available, such as a list of this type of constructions.

## 5 Conclusions

We presented a method for semantic tagging of Spanish corpora especially tailored for the needs of CPA lexicographic projects. By using Wikipedia, we expanded a previous attempt with Wiktionary, and we are now able to deal better with the large number of out-of-vocabulary units found in corpus. Integrating both approaches will probably result in better overall performance.

The general result is promising, especially considering the lightweight nature of the proposal. The method is at least robust enough for our particular purpose, and has greatly relieved the effort needed to annotate large numbers of corpus concordances, a job that otherwise would have had to be undertaken manually. We observed also that the 'conservative' variant produces high precision results, which can be useful in those scenarios where recall is not vital. This could be the case of the production of training examples

| Case | Category | Ok | Context |
|---|---|---|---|
| Bormio | Lugar | 1 | *Sus padres fueron María Zanelli y Gaetano Cerri, ambos domiciliados en el barrio Mercado Bobino y casados en* ***Bormio*** *en el año 1834* |
| enlace roto | Software | 1 | *...consultado el 17 de julio de 2011. (* ***enlace roto*** *disponible en Internet Archive; véase el historial, la primera versión y la última)* |
| General Artigas | Lugar | 1 | *En gran parte de las principales ciudades argentinas existen monumentos a Artigas o calles e incluso barrios (como "****General Artigas****", en Córdoba y, distante de tal barrio, el monumento a j. g. Artigas a la entrada del Parque Sarmiento)* |
| Amán Rawson | Lugar | 0 | *Designó ministro de gobierno a* ***Amán Rawson****, quien había militado en el partido unitario con anterioridad.* |
| asiento | Mueble | 0 | *En 1861 ingresó como soldado al batallón n.º 3 de infantería de línea con* ***asiento*** *en el fortín, hoy río cuarto, enviado a luchar contra los caudillos Ángel Vicente Peñaloza, Francisco Saá, Juan de Dios Videla y Francisco Clavero.* |
| clave | Artefacto | 0 | *Derqui mantuvo relaciones muy cordiales con el gobernador porteño, e incluso incorporó dos de los ministros de éste incluido el de hacienda,* ***clave*** *en las circunstancias que atravesaba su gobierno a su gabinete nacional* |

Table 4: Some examples of the evaluated results.

for machine learning algorithms that can associate the names of entities with elements found in their contexts of occurrence.

The study has shown many possibilities for the future. The most important is to deepen the analysis of the contexts of occurrence of the entities. In particular, we will integrate full syntactic analysis to associate semantic types with their function as argument of different verbs. Subtle cues, like syntactic relations with certain verbs, could improve the quality of the method. For instance, names referring to people tend to occupy the subject position of verbs such as *to write, smoke, argue*, etc., or in the case of the word *bombardeo* ('bombardment'), we see fragments such as the following: *el bombardeo se oía desde el palacio* ('the bombing could be heard from the palace'). The syntactic relation with the verb *oír* ('to hear') could be taken as a clear indication that the type of entity is [[Event]], because something that can be heard is an event.

Another perhaps more radical methodological change would be to go from the current text-based design (i.e., analyzing one text at a time) to a corpus-based one (analyzing all the mentions of an entity in a corpus). In this way, more information about said entity could be gathered from the corpus. Of course, with this change in perspective comes the risk of homonymy and polysemy, depending on the heterogeneity of the analyzed corpora. In a highly specialized technical corpus, however, one can expect less need for word sense induction and disambiguation mechanisms.

We also mentioned the possibility of improving the tagger, which tends to hallucinate when proposing the lemmas of the words. An idea worth exploring could be to forbid the proposal of any lemma that does not appear itself as word form in a large corpus.

Finally, we are also now replicating the method in languages such as English, French, Dutch, German and Italian. We hope the simplicity of the method will facilitate the process, but each language may pose particular challenges.

## *Acknowledgment*

This research has been made possible thanks to a grant by ANID Chile to Project Fondecyt Regular 1231594 and Project PID2022-137170OB-I00, financed by MICIU/AEI/ 10.13039/501100011033 and FEDER/UE. The authors would also like to express their gratitude to the anonymous reviewers for their outstanding work.

## *References*

Abzianidze, L. and J. Bos. 2017. Towards universal semantic tagging. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1 Long Papers*, pages 195–205.

Akbik, A., T. Bergmann, and R. Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Al-Rfou, R., V. Kulkarni, B. Perozzi, and S. Skiena, 2014. *POLYGLOT-NER: Massive Multilingual Named Entity Recognition*, pages 586–594. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Asher, N. and J. Pustejovsky. 2013. A type composition logic for generative lexicon. In J. Pustejovsky, P. Bouillon, H. Isahara, K. Kanzaki, and C. Lee, editors, *Advances in Generative Lexicon Theory*. Springer Netherlands, Dordrecht, pages 39–66.

Bird, S., E. Loper, and E. Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Bjerva, J., B. Plank, and J. Bos. 2016. Semantic tagging with deep residual networks. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541, Osaka, Japan.

Bullinaria, J. 2008. Semantic categorization using simple word co-occurrence statistics. In *ESSLLI Workshop on Distributional Lexical Semantics*.

Camacho-Collados, J. and M. T. Pilehvar. 2018. From word to sense embeddings: a survey on vector representations of meaning. *J. Artif. Int. Res.*, 63(1):743–788.

Chodorow, M., R. Byrd, and G. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting on ACL*, pages 299–304, Chicago, Illinois, USA.

Choi, E., O. Levy, Y. Choi, and L. Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia, July. Association for Computational Linguistics.

Colman, L. and C. Tiberius. 2018. A good match: a Dutch collocation, idiom and pattern dictionary combined. In J. Čibej, V. Gorjanc, I. Kosem, and S. Krek, editors, *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 233–246, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Daiber, J., M. Jakob, C. Hokamp, and P. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, pages 121–124.

Finkel, J. R., T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In K. Knight, H. T. Ng, and K. Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Gale, W. A., K. W. Church, and D. Yarowsky. 1992. One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Giacomini, L. and P. DiMuccio-Failla. 2019. Investigating semi-automatic procedures in pattern-based lexicography. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Janssen, I. Pereira,

J. Kallas, M. Jakubíček, S. Krek, and C. Tiberius, editors, *Electronic lexicography in the 21st century: Smart lexicography. Proceedings of the eLex 2019 conference*, pages 490–505. Lexical Computing.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Grishman, R. and B. M. Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, pages 466–471.

Guthrie, L., B. Slator, Y. Wilks, and R. Bruce. 1990. Is there content in empty heads? In *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90*, pages 138–143, Helsinki, Finland.

Hanks, P. 2004. The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. The MIT Press.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Hu, Z., W. Hou, and X. Liu. 2024. Deep learning for named entity recognition: a survey. *Neural Computing & Applications*, 36:8995–9022.

Huang, J. Y., B. Li, J. Xu, and M. Chen. 2022. Unified semantic typing with meaningful label inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654. Association for Computational Linguistics, July.

Ježek, E., B. Magnini, A. Feltracco, A. Bianchini, and O. Popescu. 2014. T-pas; a resource of typed predicate argument structures for linguistic analysis and semantic processing. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895. European Language Resources Association (ELRA).

Kutbi, M. 2023. Named entity recognition utilized to enhance text classification while preserving privacy. *IEEE Access*, 11:117576–117581.

Lesk, M. E. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Li, J., A. Sun, J. Han, and C. Li. 2023. A survey on deep learning for named entity recognition. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3817–3818.

Light, M. 1996. Morphological cues for lexical semantics. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, page 25–31, USA. Association for Computational Linguistics.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 2, COLING '98*, pages 768–774.

Ling, X. and D. S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1038–1047.

Marini, C. and E. Ježek. 2019. CROATPAS: A resource of corpus-derived typed predicate argument structures for Croatian. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Meyer, I. 2001. Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, and M. L'Homme, editors, *Recent Advances in Computational Terminology*. John Benjamins, pages 279–302.

Moro, A., F. Cecconi, and R. Navigli. 2014. Multilingual word sense disambiguation

and entity linking for everybody. In *Proceedings of the 13th International Semantic Web Conference, Posters and Demonstrations (ISWC 2014)*, pages 25–28, Riva del Garda, Italy.

Nadeau, D. 2007. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. Ph.D. thesis, Department of Information Technology and Engineering, University of Ottawa.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(7).

Namer, F. 2002. Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. In J.-M. Pierrel, editor, *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 237–246, Nancy, France. ATALA.

Navigli, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41:10:1–10:69.

Nazar, R. forthcoming. Semantic typing for corpus pattern analysis. *International Journal of Lexicography*.

Panchenko, A., S. Faralli, E. Ruppert, S. Remus, H. Naets, C. Fairon, S. Ponzetto, and C. Biemann. 2016. Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327.

Pearson, J. 1998. *Terms in context*. John Benjamins, Amsterdam.

Piao, S., P. Rayson, D. Archer, F. Bianchi, C. Dayrell, M. El-Haj, R.-M. Jiménez, D. Knight, M. Kren, L. Löfberg, et al. 2016. Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016)*, pages 2614–2619.

Pustejovsky, J., P. Hanks, and A. Rumshisky. 2004. Automated induction of sense in context. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*.

Rayson, P., D. Archer, S. Piao, and T. McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, in association with LREC-04*, pages 7–12. European Language Resources Association.

Renau, I., R. Nazar, A. Castro, B. López, and J. Obreque. 2019. Verbo y contexto de uso: Un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos*, 52(101):878–901.

Renau, I., R. Nazar, and D. M. Melanchthon. 2024. Towards the automatic generation of a pattern-based dictionary of Spanish verbs. In K. Despot, A. Ostroški Anić, and I. Brač, editors, *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, pages 367–383, Cavtat. Institut za hrvatski jezik.

Sarkar, R., J. McCrae, and P. Buitelaar. 2018. A supervised approach to taxonomy extraction using word embeddings. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2059–2064.

Sekine, S. and C. Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *Language Resources and Evaluation Conference (LREC)*, pages 1977–1980.

Shwartz, V., E. Santus, and D. Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1*, pages 65–75.

Straka, M. and J. Straková. 2018. UDPipe: Universal dependency parser. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, pages 157–167.

Tjong Kim Sang, E. F. and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume*

*4*, CONLL '03, page 142–147, USA. Association for Computational Linguistics.

Weeds, J. and D. Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 81–88.

Wiedemann, G., S. Remus, A. Chawla, and C. Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *CoRR*, abs/1909.10430.

Wilks, Y. and M. Stevenson. 1997. Sense tagging: Semantic tagging with a lexicon. In *Tagging Text with Lexical Semantics: Why, What, and How?*, pages 47–51. Association for Computational Linguistics.

Xiong, W., J. Wu, D. Lei, M. Yu, S. Chang, X. Guo, and W. Y. Wang. 2019. Imposing label-relational inductive bias for extremely fine-grained entity typing. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 773–784, Minneapolis, Minnesota. Association for Computational Linguistics.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, page 454–460. Association for Computational Linguistics.