Increasing accuracy of a semantic word labelling tool based on a small lexicon

Hugo Sanjurjo-González
University of Deusto, Spain
hugo.sanjurjo@deusto.es

Abstract

Semantic annotation has become an important piece of information within corpus linguistics. This information is usually included for every lexical unit of the corpus providing a more exhaustive analysis of language. There are some resources such as lexicons or ontologies that allow this type of annotation. However, expanding these resources is a time-consuming task. This paper describes a simple NLP baseline for increasing accuracy of the existing semantic resources of the UCREL Semantic Analysis System (USAS). In our experiments, Spanish token accuracy is improved by up to 30% using this method.

1 Introduction

Apart from raw texts, a corpus can include extra linguistic information by way of annotation. Most common types of annotation are grammatical, semantic, prosodic and historical. The semantic one has become an important piece of information within the corpus linguistics research field. A corpus with this information is a useful resource to extract knowledge from a real context: as Navarro et al. (2005) state, it can be considered as a semi-structured database that offers deep information about human knowledge, concepts and relations among them.

Semantic annotation in corpus linguistics tends to recognise semantic categories and concepts at different syntactic levels, such as word level, phrase level or sentence level (Piao et al., 2018). For this purpose, the information about grammatical tags and NER (Named-Entity Recognition) classes contribute to determine lexical semantics to some extent, but they are not sufficiently informative (Abzianidze and Bos, 2017). Semantic annotation tries to overcome these barriers by adding new categories.

In this paper, we describe an NLP baseline that increases accuracy of a semantic role labelling tool that makes use of a small semantic lexicon in Spanish language (Piao et al., 2015) based on the USAS tagset (Archer et al., 2002). We are able to increase accuracy by means of a very simple strategy that makes use of freely-available NLP toolkits such as NLTK (Bird et al., 2009) and Spacy (Honnibal and Montani, 2017). A novel approach using WordNet similarity based on the information content theory (Resnik, 1995) is also employed in order to search synonyms of unknown words and, therefore, increase lexical accuracy. As a proof of concept, we carried out different experiments with texts from the finance domain.

The article is organised as follows: Section 2 explains our approach together with the different processes that are executed. Implementation is described in Section 3. We show different experiments in Section 4. Last, Section 5 outlines conclusions and future directions.

2 Overview of our approach

The USAS lexicon is based on the Longman Lexicon of Contemporary English (McArthur, 1986), which ensures, up to a certain point, the linguistic validity and motivation of this resource. There are 21 major discourse fields, expanding into 232 category labels[1]. USAS employs a group of labels in an attempt to include most meanings of the lexicalised unit. To employ USAS lexicon, we need to extract lemma and grammatical annotation for each word. Table 1 shows an example of entry for the word *business*.

Regarding the Spanish version, it contains around 10,000 words and 5,000 multiword expressions, and most of them are Spanish named entities such as places or locations. As a consequence of its

[1] More information can be found in (Archer et al., 2002)

Proceedings of the 17th International Conference on Natural Language Processing, pages 10–14
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

Table 1: Example of lexicon entry.

Lemma	POS	Semantic tags
business	noun	I2.1 A1.1.1 A5.1+++

reduced size, accuracy of the Spanish USAS lexicon is limited if it is used in specific text domains. For instance, only 3.30% of the lexicon entries belong to the finance domain. If this is the only resource employed for tagging, there will be many words that will not have any tag, and thus, many words will be incorrectly tagged as unknown because they do not appear in the lexicon (e.g. *índice* - index).

A more in-depth analysis reveals that some of these words are lemmatised incorrectly (e.g. *véase* (note) is lemmatised as *véase* instead of *ver*). In addition, some words appear in the lexicon with only one grammatical category when they can belong to different categories (e.g. *mucho* (many/much) can be an adverb and adjective in Spanish).

To solve these problems, the simplest solution is to add new entries to the lexicon, however this is a very time-consuming task. Another solution is trying to improve results of the operations required by USAS such as lemmatisation or grammatical annotation. We can also try to incorporate other techniques such as stemming in order to match the stem of the word with another stem in the lexicon. In order to achieve that, we can simply employ available Spanish resources from NLP toolkits such as Spacy and NLTK. In the rest of this section, we will describe how lexicon accuracy may be increased using some preprocessing techniques in a specific domain such as the finance one.

2.1 Analysing finance domain texts

In this stage, we built a corpus of texts from the finance domain in order to analyse its main features such as most frequent words, keywords, collocations etc. More concretely we selected the Annual Report of the Banco de España (1998-2019) (BDE, 2020) with the exception of the 2013 edition, since it was used for validation purposes. These documents review economic and financial developments in the Spanish economy and are composed of 19 samples and 2,841,826 words.

Analysis of this corpus reveals that this type of texts appear to have many acronyms such as PIB (*Producto Interior Bruto* - Gross Domestic Product), numbers with different formats (2005, 36,3%, 540.000, 3,25), currency symbols, proper names (Miguel Fernández Ordóñez Antonio Rosas), geographical names (Torrellano-Elche, *Eurozona*), organisations names (*Banco de España*) and words in languages different from Spanish (financial institutions) as well as other jargon of this domain.

2.2 Lemmatisation

As we previously mentioned, USAS lexicon entries are composed of lemma, POS tag and semantic tag (see Table 1). Thus, it is necessary to include a Spanish lemmatiser. NLTK does not offer this tool for Spanish language, and Spacy includes one but it has some errors. For instance: *reclamaciones* as *reclamaciones* (claims in English) or *como* (like) as *comer* (eat), para (for) as *parir* (give birth), among others.

Using full words instead of lemmas also entails some errors because most of them do not appear in the lexicon. For instance, *músculos* - muscles instead of *músculo* - muscle. For this reason, we make use of the NLTK stemmer, which returns words' bases or roots.

2.3 Grammatical annotation

We also need to annotate each word grammatically. English USAS employs CLAWS (Garside, 1987), a highly sophisticated grammatical tagger. However, there is not an equivalent for other languages, so in this case USAS for Spanish employs a simplified version of the grammatical tagset that includes the basic grammatical elements.

For this purpose, we employ Spacy grammatical tagger since it offers a relatively adequate performance. Nevertheless, some words are incorrectly tagged, mainly some nouns or even adjectives that were tagged as proper nouns because of their initial uppercase. For instance: *Informe* (Report) and *Anual* (Annual). As a consequence, the semantic tagger would return no tag for all these words. To overcome this problem we search for words without grammatical tags in the lexicon at the end of the process, that is, as a final measure to return semantic tag candidates.

2.4 Identifying named entities and foreign words

We make use of a corpus of names included in NLTK (Kantrowitz, 2020), for instance Alberto. This allows us to identify any name in different languages. We also employ some gazetteers for geographical locations (e.g. Madrid) and the previously mentioned corpus for identifying English words that usually appear in financial texts (e.g. Exchange).

Including a NER tagger for Spanish is also an option, however according to our experiments this tool recognises many foreign words such as organisations or even locations (e.g. Cash). For this reason, if we included it, it would return many false positives.

2.5 Identifying other elements

In order to identify any format number, mathematical operations and symbols as well as some other elements like abbreviations, we formulate patterns using Perl compatible regular expressions.

2.6 Computing WordNet synonyms

We also wanted to make use of a novel approach for identifying unknown words that were not tagged in previous steps. To do that, we try to get synonyms of the unknown words, since synonyms often have the same semantic function.

We employ sense similarity of the information content of the corpus compiled at the first stage of this approach. Our premise is that if one word is missing from the lexicon there are many possibilities that this word has a synonym in the previously compiled corpus.

We create an information content dictionary of the corpus in order to employ similarity based on the WordNet synsets. To measure similarity we employ Lin measure (Lin, 1998).

3 Implementation and deployment

We develop all the components of the tagger following the specifications proposed in the previous section. Fig. 1 shows a simplified workflow of this process that can be described as follows:

1. First, the tagger searches if the lemma of the word together with its grammatical tag is in the lexicon. If it is, we already have the tag for the word.

2. If it is not, the tagger searches if the word with its grammatical tag is in the lexicon.

3. If not, we employ the stemmed version of the word and the lexicon together with the grammatical tag.

4. If we do not have any results, we try the same without using grammatical annotation as a consequence of the possible errors of the grammatical tagger.

5. After that, words without semantic tags are analysed in order to identify named entities and foreign words.

6. Subsequently, regular expressions are used to match numbers and abbreviations, among others.

7. For the rest of the unmatched words, the tagger will search a synonym of the word using the information content and its similarity based on WordNet synsets. We get a list of candidates according to their similarity index and search them in the lexicon.

8. If similar words cannot be calculated with that word or its lemma, it will be set as a semantic tag ‘Z99’ or unknown word.

Table 2: Evaluation of accuracy.

Sample text size	Correct	Partially correct
13,331 words	86.26%	2.21%
7,064 words	86.71%	1.33%

4 Experiments

In the absence of resources for validating our tool we needed to build a custom-made gold standard. This is a consequence of the USAS tagset, a very specific classification system, and the Spanish language, which has less lexical coverage than English language using this resource. We extracted some sample texts that were not included in the corpus together with some texts from independent sources. The size of the gold standard is 20,395 words. This size is a consequence of the laborious task of manual annotation.

In order to evaluate the accuracy, we followed the same metrics as (Piao et al., 2015). A first metric refers to those instances where the first candidate tag is correct, and a second metric makes reference to the cases where the other tags in the list are correct or closely related to the word sense. These results are shown in Table 2

As we can see in Fig. 2, results have been improved around 30% in comparison with using only Lemma – POS method.

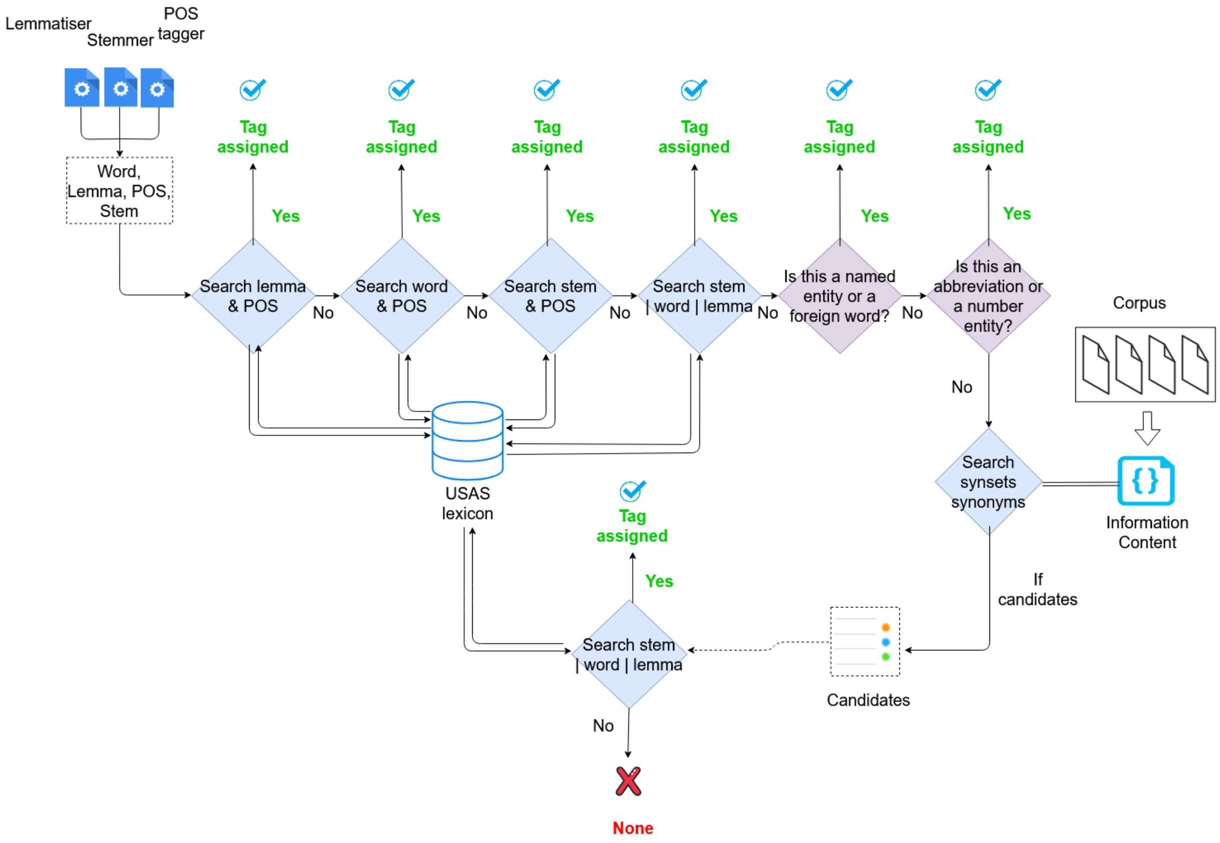

Figure 1: Simplified workflow diagram of the tagger.

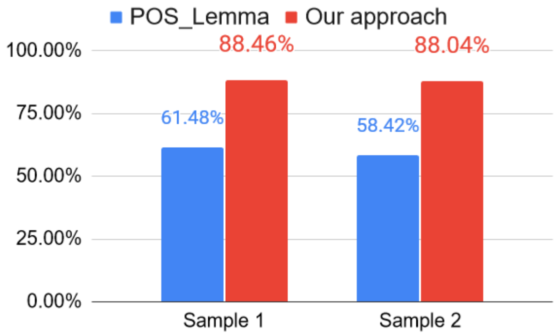

Figure 2: Experiment results.

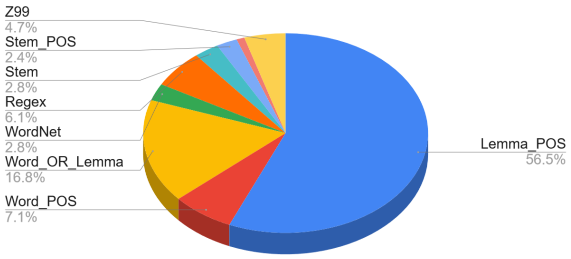

Figure 3: % of words that were tagged for each subprocess.

Fig. 3 shows percentages of words that were tagged for each subprocess of the tagger. As it can be seen, the proposed baseline tags about 44% of the words. WordNet synonym method did not return any significant results, maybe as a consequence of the absence of a basis of finance elements in the lexicon. Its inclusion only improves accuracy around 0.15% according to our experiments, so it is not significant.

Last, the confusion matrix of the semantic tagger according to the 21 major discourse fields of the USAS taxonomy can be found in the Supplementary Material. Confusion matrix of all the 232 subcategories would be a more detailed option but the representation of all the subcategories may be slightly confusing. The most remarkable issues are the following:

- Many words from the rest of the categories are incorrectly tagged using Z category. More specifically with 'Z99' tag. That means that there are many words that our tagger cannot recognised and as a consequence they are tagged as unknown.

- Another issue is related to the words belonging to the N category (Numbers and Measurement) that are wrongly tagged using the A category (General and Abstract terms). Words such as *índice* (index) or *tasa* (rate, fee) are not correctly identified.

- Last, it should be mentioned that words belonging to A category are tagged incorrectly using the rest of the categories. One explanation may be the own definition of this category, general and abstract terms.

5 Conclusions and further work

The main contribution of this paper is a strategy that utilises existing NLP toolkits such as NLTK and Spacy to preprocess texts in order to obtain better results using only a small lexicon as source of semantic information. This strategy is implemented following a simple and straightforward approach. Empirical results are reported and compared across an ad hoc gold standard based on texts from the finance domain.

This study also introduced a novel approach for extending lexical accuracy of semantic lexicons by means of synsets similarity of WordNet which did not provide the expected results, maybe as a consequence of the limited lexicon.

We hope that this approach could be easily extended to other domains and even with under-resourced languages. Therefore, expected future work includes reproducing the good results obtained in other text domains and employing languages different from Spanish or English. We also need to investigate how to take advantage of the semantic similarity provided by WordNet or even word embeddings using taxonomies like USAS.

References

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Dawn Archer, Andrew Wilson, and Paul Rayson. 2002. Introduction to the usas category system. retrieved from ucrel semantic analysis system (usas) website. http://ucrel.lancs.ac.uk/usas/usas_guide.pdf. Accessed: 2020-09-16.

BDE. 2020. Banco de españa - publicaciones anuales website. https://www.bde.es/bde/es/secciones/informes/Publicaciones_an/Informe_anual/. Accessed: 2020-09-16.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Roger Garside. 1987. The claws word-tagging system. *The Computational analysis of English: A corpus-based approach. London: Longman*, pages 30–41.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Mark Kantrowitz. 2020. Mark kantrowitz corpora names website. http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/. Accessed: 2020-09-16.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Machine Learningˆ* Proceedings of the Fifteenth International Conference (ICML'98)*, pages 296–304.

Tom McArthur. 1986. *Longman lexicon of contemporary English*. Longman.

Borja Navarro, Patricio Martínez-Barco, and Manuel Palomar. 2005. Semantic annotation of a natural language corpus for knowledge extraction. In *International Conference on Application of Natural Language to Information Systems*, pages 365–368. Springer.

Scott SL Piao, Francesca Bianchi, Carmen Dayrell, Angela D'egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1274.

Scott SL Piao, Paul Rayson, Dawn Knight, and Gareth Watkins. 2018. Towards a welsh semantic annotation system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

P Resnik. 1995. Using information content to evaluate semantic similarity. In *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada*, pages 448–453.