Open Information Extraction for Spanish Language based on Syntactic Constraints

Alisa Zhila
Centro de Investigación en Computación,
Instituto Politécnico Nacional,
07738, Mexico City, Mexico
alisa.zhila@gmail.com

Alexander Gelbukh
Centro de Investigación en Computación,
Instituto Politécnico Nacional,
07738, Mexico City, Mexico
gelbukh@gelbukh.com

Abstract

Open Information Extraction (Open IE) serves for the analysis of vast amounts of texts by extraction of assertions, or relations, in the form of tuples ⟨*argument 1*; *relation*; *argument 2*⟩. Various approaches to Open IE have been designed to perform in a fast, unsupervised manner. All of them require language specific information for their implementation. In this work, we introduce an approach to Open IE based on syntactic constraints over POS tag sequences targeted at Spanish language. We describe the rules specific for Spanish language constructions and their implementation in EXTRHECH, an Open IE system for Spanish. We also discuss language-specific issues of implementation. We compare EXTRHECH's performance with that of REVERB, a similar Open IE system for English, on a parallel dataset and show that these systems perform at a very similar level. We also compare EXTRHECH's performance on a dataset of grammatically correct sentences against its performance on a dataset of random texts extracted from the Web, drastically different in their quality from the first dataset. The latter experiment shows robustness of EXTRHECH on texts from the Web.

1 Introduction

Open IE is a rapidly developing area in text processing, with its own applications and approaches that are different from traditional IE (Etzioni et al., 2008; Banko and Etzioni, 2008; Etzioni, 2011). Unlike traditional IE, where systems are targeted at extraction of instances of particular relations with arguments restricted to certain semantic classes, e.g., *to_be_born_in(HUMAN; LOCATION)*, Open IE serves for extraction of all possible relations with arbitrary arguments. For example, in "*Woman who drove van full of kids is charged with attempted murder*" two relations can be identified: ⟨*Woman*; *drove*; *van full of kids*⟩ and ⟨*Woman*; *is charged with*; *attempted murder*⟩.

The ability to extract arbitrary relations from text allows applications of Open IE that are not possible in the frame of traditional IE. Among them are fact extraction at sentence level (e.g., ⟨*Mozart*; *was born in*; *Salzburg*⟩), new perspective on search as question answering (e.g., *Where was Mozart born?*) (Etzioni, 2011), or assessment of the quality of text documents at Web scale (Horn et al., 2013). Additionally, the output of Open IE systems can serve for ontology population (Soderland et al., 2010) and acquisition of common sense knowledge (Lin et al., 2010).

Although all Open IE systems are targeted at the extraction of arbitrary relations, the approaches to this task vary significantly. The pilot approach suggested by Banko et al. (2007) is based on semi-supervised learning of general relation patterns that then serve for extraction of arbitrary relations. However, the output of such systems contains many incoherent and inconsistent extractions, and the training stage is quite computationally complex. Fader et al. (2011) suggested another approach where syntactic and lexical constraints were applied over POS-tagged input. This approach has proven to be robust and fast enough for relation extraction at Web scale.

Although Open IE is targeted at extraction of arbitrary relations without any semantic restrictions, all approaches have strong language dependent restrictions and require language specific information to be introduced in the corresponding systems. For Spanish language, the apporach based on rules over dependency trees has been implemented both using full parsing

Proceedings of the ACL 2014 Student Research Workshop, pages 78–85,
Baltimore, Maryland USA, June 22-27 2014.

(Aguilar-Galicia, 2012) and using shallow dependency parsing (Gamallo et al., 2012). The former work shows that this approach is too computationally costly and is not always robust even on grammatically correct texts. The latter work does not report any results for Spanish language or discusses any details specific to implementations for languages other than English. Further, we are not aware of any existing research on whether the approach based on syntactic constraints over POS tags can be generalized to other languages. Additionally, although Open IE is claimed to be useful for information extraction from the Web, we are not aware of any research on its applicability to texts randomly extracted from the Internet, i.e., those that have not been verified for grammatical correctness by peers or editors.

In this paper we discuss Open IE based on syntactic constraints over POS tag sequences, aimed at Spanish language. We describe its implementation and introduce EXTRHECH, an Open IE system for Spanish. We also compare its performance with that of REVERB (Fader et al., 2011) on a parallel dataset. Additionally, we evaluate performance of our system over a dataset of texts randomly extracted from the Internet and discuss the issues that arise when processing random Internet texts. We also give a brief analysis of errors.

The paper is organized as follows. Related work is reviewed in Section 2. Section 3 presents our approach to Open IE for Spanish and describes the EXTRHECH system. Section 4 describes the experiments for a parallel English-Spanish dataset and for a Spanish dataset of texts randomly extracted from the Internet. In Section 5, a brief analysis of errors is presented. Section 6 draws the conclusions and outlines future work.

2 Related Work

There exist several approaches to Open IE.

Chronologically the first one was introduced in the pilot works on Open IE by Banko et al. (2007) and Etzioni et al. (2008). Their approach is based on semi-supervised machine learning principles and includes three main steps: (1) manual labeling of a training corpus for seed relation phrases and features; (2) further semi-supervised learning of relations; (3) automatic extractions of relations and their arguments. This approach is implemented in TEXTRUNNER (Banko and Etzioni, 2008), WOE^{pos}, and WOE^{parse}, both (Wu and Weld, 2010). In these systems, the detection of a relation triple starts from the potential arguments expressed as noun phrases, i.e., before the connecting relation phrase is detected. Once detected, neither the argument phrases nor the relation phrase can be backtracked, which makes the approach prone to incoherent and uninformative extractions. For example, in "*to make a deal with*", *deal* can be erroneously extracted as an argument, although it is a part of the relation phrase.

The group of rule-based approaches includes systems based on rules applied over linguistically annotated texts. FES-2012 system (Aguilar-Galicia, 2012) applies rules to the fully parsed sentences. However, in the same work the authors show that this approach is too slow to be scaled to a Web-sized corpus and that it is not robust. Another system implementing rule-based approach is DEPOE (Gamallo et al., 2012). In this system, the rules are applied to the output of shallow dependency parsing. In REVERB system (Fader et al., 2011), syntactic constraints are applied over POS tags and syntactic chunks. The last two systems show better results in terms of precision/recall and speed, and, consequently, scalability to a Web-sized corpus.

Finally, the approach based on the deep automatic linguistic analysis is implemented in OLLIE (Mausam et al., 2012). This system combines various approaches: it uses output of a rule-based Open IE system to bootstrap learning of the relation patterns and then additionally applies lexical and semantic patterns to extract relations that are not expessed through verb phrases. Such a complex approach leads to high-precision results with a high yield. However, there is a tradeoff between accuracy of the output and cost of implementation and computation and complexity of the training stage.

All these approaches require language-dependent information for their implementation. The third approach directly uses lexical information for the context analysis. The other two approaches employ language-specific morphological and syntactic information. Of the described systems, only two have been implemented for languages other than English. FES-2012 system is implemented for Spanish language; however, its use of the full syntactic parsing does not scale to a Web-sized corpus. DEPOE system, based on rules over shallow dependency parsing, is claimed

to have its variants for Spanish, Portuguese, and Galician languages (Gamallo et al., 2012). However, the authors do not report any experimental results on languages other than English or any language-specific details.

The approach based on syntactic constraints over POS tags has not been applied to languages other than English, in spite of that this method can be easily adapted to other languages because it only requires a reliable POS tagger. The basic algorithm for relation extraction, according to Fader et al. (2011), is as follows:

- First, search for a verb-containing relation phrase in a sentence;
- If detected, search for a noun phrase to the left of the relation phrase;
- If a noun phrase detected, search for another noun phrase to the right of the relation phrase.

Additionally, the experiments for Open IE systems have been conducted only on texts that came from verified sources, i.e., Wikipedia, news, or textbooks (Banko and Etzioni, 2008; Fader et al., 2011; Mausam et al., 2012). However, Open IE is meant to work with Web text data that may come from any source including those that have not been edited or verified for grammar errors.

3 System Description

In this section we introduce EXTRHECH,[1] a system for Open IE in Spanish. It takes a POS-tagged text as input, applies syntactic constraints over sequences of POS-tags, and returns a list of extracted relations as triples ⟨*argument 1*; *relation*; *argument 2*⟩ that correspond to each sentence.

3.1 Basic Processing

The system takes as input a POS-tagged text. In our experiemnts, we used a morphological analyzer from Freeling-2.2 (Padró et al., 2010). For Spanish language, it returns POS tags accoridng to EAGLES POS tag set (Leech and Wilson, 1999). Consequently, our system is designed to work with this POS tag set.

Spanish uses a number of non-ASCII characters, such as *á, é, ñ*, etc. These characters can come in different encodings. To be able to correctly analyze text with these characters, Freeling analyzer should receive the input in ISO encoding. Thus, the input text needs an additional preprocessing stage to be converted into this encoding. Though this might look as a minor technical issue, guessing the original encoding becomes a significant problem when working with texts from arbitrary sources on the Web. We discuss encoding related issues in Section 4.2.

After the text has been properly POS-tagged, we feed it into EXTRHECH system, which applies the fact extraction algorithm described in Section 2 to each sentence, one sentence at a time. We use the same basic algorithm as in (Fader et al., 2011) but with different triple matching rules as appropriate for Spanish grammar.

The original POS-tag sequences for English would produce nonsense results on Spanish input due to substantial difference in grammars: infinitives are not preceded by "*to*", adjectives usually follow nouns, and oblique case pronouns precede verbs instead of following them, just to name a few peculiarities of Spanish.

First, the system looks for a verb-containing phrase in a sentence by matching it against the following expression:

$$\mathsf{VREL} \rightarrow (\mathsf{V}\ \mathsf{W}^*\ \mathsf{P}) \mid (\mathsf{V}),$$

where V stands either for a single verb optionally preceded by a reflexive pronoun (*se realizaron*, "*were carried out*"), or a participle (*calificado*, "*qualified*"). V W* P matches a verb with dependent words, where W stands for either a noun, an adjective, an adverb, a pronoun, or an article, and P stands either for a preposition optionally immediately followed by an infinitive, or for a gerund (*sigue siendo*, "*continues to be*"). The symbol * denotes zero or more matches. Here and further, the whole match is referred to as *verb phrase* (though it is not a verb phrase in linguistic sense).

After detecting a verb phrase, EXTRHECH looks for a noun phrase to the left from the beginning of the verb phrase. This noun phrase is a potential first argument of the relation. If a match is found, then the system looks for another noun phrase to the right from the end of the verb phrase. The noun on the right side is treated as the second argument.

Noun phrases are searched for with the following regular expression:

$$\mathsf{NP} \rightarrow \mathsf{Np}\ (\mathsf{PREP}\ \mathsf{Np})?,$$

where Np matches a noun optionally preceded by either an article (*la dinámica*, "*the dynamics*"),

[1] All materials are available on the page http://www.gelbukh.com/resources/spanish-open-fact-extraction.

an adjective, an ordinal number (*los primeros ganadores*, "*the first winners*"), a number (*3 casas*, "*3 houses*"), or their combination, and optionally followed by either a single adjective (*un esfuerzo criminal*, "*a criminal effort*"), a single participle, or both (*los documentos escritos antiguos*, "*the ancient written documents*"). The whole expression matched by Np can be preceded by an indefinite determinant construction, e.g., *uno de*, "*one of*". PREP matches a single preposition. Hence, an entire noun phrase is either a single noun with optional modifiers or a noun with optional modifiers followed by a prepositional phrase that is a preposition and another noun with its corresponding optional modifiers (*una larga lista de problemas actuales*, "*a long list of current problems*"). The symbol ? denotes 0 or 1 matches.

If noun phrases are matched on both sides of the verb phrase, all three components are considered to represent a relation and are extracted in the form of a triple.

As an output unit, EXTRHECH returns a triple consisting of ⟨*argument 1*; *relation*; *argument 2*⟩, where *argument 1* semantically is, e.g., an agent or experiencer of the relation and *argument 2* is a general object or circumstance of the relation.

3.2 Additional Processing

Above we described the core rules and the basic sequence for relation extraction. In addition to them, we also implemented several optional rules for processing of certain language constructions that can be turned on and off with the input parameters.

First, participle clauses that follow a noun can be searched for a relational triple if they terminate with a noun. For example, from a phrase

Precios del café suministrados por la OIC
("Coffee prices provided by International Coffee Organization")

EXTRHECH returns the relation:

⟨*Precios del café*; *suministrados por*; *la OIC*⟩.

Second, EXTRHECH also approaches resolution of coordinating conjunctions between verb phrases and between noun phrases into corresponding separate relations. Here follows the example of a sentence with a coordinating conjunction between verb phrases:

El cerebro almacena enormes cantidades de información y realiza millones de actividades todos los días
("The brain stores vast amounts of information and performs millions of activities every day")

. Two facts are detected:

⟨*El cerebro*; *almacena enormes cantidades de*; *información*⟩
and
⟨*El cerebro*; *realiza millones de*; *actividades todos los días*⟩.

Third, relative clauses introduced by single relative pronouns (e.g., *que* (*"that", "who"*), *cual* ("*which*")) as in *las partes que conforman un trabajo de investigación* ("*parts that make up a research work*") are also searched for relations. However, relative pronoun phrases with prepositions, e.g. *en el cual* ("*in which*") are not taken into consideration for relation extraction due to their coreferential complexity.

3.3 Limitations

The implementation of basic processing performed by EXTRHECH system follows the algorithm introduced in (Fader et al., 2011). This means that extracted facts are limited to the relations expressed through a verb phrase. This limitation is discussed in the cited paper.

In our apporach to Open IE in Spanish, we do not allow pronouns to be potential arguments of a relation. It was mainly done because of a wide use of a neutral pronoun *lo* ("*this*", "*which*", or no direct translation) as a head of relative clauses in Spanish language, e.g., *lo que dio valor al poder judicial* ("*__ that gave value to the judiciary*"). Including pronouns for potential argument matches would return a lot of uninformative relations as ⟨*lo*; *dio valor a*; *el poder judicial*⟩. This issue can be solved only by introducing anaphora resolution techniques which involves processing on a supersentence level. Although seemingly feasible, this modification will necessarily slow down the extraction speed which is critical while working with large scale corpora. As mentioned in Section 2, high speed performance is one of the main advantages of the approach to Open IE based on syntactis constraints compared to the others. Hence, any modifications that would affect its speed should be considered with caution.

Another language dependent limitation is related to the order of the processing. As earlier described in Section 3.1, an extracted triple is expected to correspond semantically to ⟨*agent/experiencer*; *relation*; *general object/circumstance*⟩. This is expected to be correct for a direct word order, i.e., Subject – Verb – (Indirect) Object, which is a dominant word order for Spanish. Yet the inverted word order, i.e.

(Indirect) Object – Verb – Subject (e.g., *De la médula espinal nacen los nervios periféricos*, i.e., literally *"*From the spinal cord arise peripheral nerves*"), also occasionally takes place in grammatically correct and stylistically neutral Spanish texts. However, the occurence of this construction is less then 10% according to (Clements, 2006).

4 Experiments and Evaluation

In this section we describe the experiments conducted with EXTRHECH system.

4.1 Experiment on parallel news dataset

We compare EXTRHECH's performance with that of REVERB, an Open IE system for English based on the same algorithm (Fader et al., 2011). Since these systems are designed for different languages, we ran our experiment on a parallel dataset.[1]

We took 300 parallel sentences from the English-Spanish part of News Commentary Corpus (Callison-Burch et al., 2011). Then, we ran the extractors over the corresponding languages. After that, two human annotators labeled each extraction as correct or incorrect. For the Spanish part of the dataset, the annotators agreed on 80% of extractions (Cohen's kappa $\kappa = 0.60$), whereas for the English part they agreed on 85% of extractions with $\kappa = 0.68$. For both datasets their respective κ coefficients indicate substantial agreement between the annotators.

Precision was calculated as a fraction of correct extractions among all returned extractions. We calculated *Recall* as a fraction of all returned correct extractions among all possible (i.e., expected) correct extractions. By manual revision of the sentences in the datasets, we made a list of all expected correct extractions. Their number was used to estimate the recall.

In contrast to REVERB, our system does not have a confidence score mechnaism at this point. To make the comparison between the systems appropriate, we ran REVERB extractor with the confidence score level set to 0 that means that the system returns all relations that match the rules, i.e., in the same way as EXTRHECH does. Hence, the systems were in equivalent conditions. The results of the experiment are shown in Table 1.

As we see, on a parallel dataset of texts from News Commentary Corpus, both systems show a very similar performance. Based on this observation, we can conclude that the algorithm suggested in (Fader et al., 2011) can be easily adopted for other languages with dominating SVO word order and an available POS-tagger.

System	Precision	Recall	Correct Extractions	Returned Extractions
EXTRHECH	0.59	0.48	218	368
REVERB	0.56	0.44	201	358

Table 1: Performance comparison of REVERB and EXTRHECH systems over a parallel dataset.

4.2 Experiment on Raw Web dataset

One of the most important goals of Open IE systems is to be able to process large amounts of texts directly from the Web. This requires high performance speed and robustness on texts that often lack grammatical and orthographical correctness or coherence. The study showing the approach's advantage in speed was already presented in (Fader et al., 2011). In this work we focused on robustness. We evaluated the performance of our system on a dataset of sentences extracted from the Internet "as is". For this dataset, we took 200 random data chunks detected by a sentence splitter from CommonCrawl 2012 corpus (Kirkpatrick, 2011), which is a collection of web texts crawled from over 5 billion web pages. However, 41 from those 200 chunks were not samples of textual information in human language but rather pieces of programming codes or numbers. We took out these chunks because they are not relevant for our research. In a real life scenario they could be easily detected and eliminated from the Web data stream. After this, our dataset consisted of 159 sentences written in human language. We will refer to this dataset as Raw Web text dataset.[1] Of 159 sentences of the dataset, 36 sentences (22% of the dataset) were grammatically incorrect or incoherent, as evaluated by a professional linguist.

We ran EXTRHECH system over this dataset and asked two human judges to label extractions as correct or incorrect. The annotators agreed on 70% of extractions with Cohen's $\kappa = 0.40$, which indicates the lower bound of moderate agreement between judges.

Precision and *Recall* were calculated in the same manner as described in Section 4.1. We compare these numbers to the results obtained for the dataset of grammatically correct sentences from News Commentary Corpus in Table 2.

We can observe that system's performence has

Dataset	Precision	Recall
News Commentary	0.59	0.48
Raw Web	0.55	0.49

Table 2: Performance of EXTRHECH on the grammatically correct dataset and the dataset of noisy sentences extracted from the Web

not lowered significantly when processing "noisy" texts compared to edited newspaper texts. An interesting observation is that texts from the Internet are poorer in facts than the news texts. The number of expected extractions was manually evaluated by a human expert for both datasets. The ratio of extractions to sentences for the news dataset was 1.5:1, while for the Raw Web dataset it was only 1.03:1.

Now we will briefly discuss the issue arising due to various encoding standards used for non-ASCII characters, e.g., of *á, é, ñ*, etc. While applying Freeling morphological analyzer to the dataset, we encountered an issue that the sentences came in various encodings. As we mentioned in Section 3, Freeling-2.2 analyzer works properly only with ISO encoded input. Therefore, we had to convert each sentence from the dataset into ISO encoding. While most of the sentences were in UTF-8 encoding and were converted in a single pass, the encoding of about 3% of the sentences was initially corrupted, therefore, they were not processed correctly by the POS-tagger. Although the issue is manageable at the scale of a small dataset, it might affect the speed and quality of fact extraction when working at Web scale.

5 Error Analysis

After running EXTRHECH on the datasets, we analyzied the errors in the output. We followed the classifications of the types of errors and their causes suggested in (Zhila and Gelbukh, 2014). The distribution of the errors in EXTRHECH's output over the types of errors is shown in Table 3. The data about error types was gathered over extractions from Raw Web dataset. When errors are present both in the arguments and in the relation phrase, they are likely to have the same cause.

Based on the analysis of the outputs over Raw Web dataset, the following causes for errors have been observed:

- Underspecified noun phrase
- Overspecified verb phrase
- Non-contiguous verb phrase

Type of errors	Percentage
Incorrect relation phrase	21%
Incorrect argument(s)	45%
of them, with also incorrect relation	19%
Incorrect argument order	6%

Table 3: Distribution of errors in output by the basic error types in relation extraction for EXTRHECH system run over Raw Web dataset

- N-ary relation or preposition (e.g., *entre*, "*between*")
- Conditional subordinate clause
- Incorrectly resolved relative clause
- Incorrectly resolved conjunction
- Inverse word order
- Incorrect POS-tagging
- Grammatical errors in original sentences

Inverse word order is one of the main causes for the incorrect order of arguments in extracted relations. However, as it can be seen in Table 3, this is the least common type of errors, which is in accordance to the low frecuency of the inverse word order (Clements, 2006). A more detailed analysis of the issues that cause the errors can be found in (Zhila and Gelbukh, 2014).

6 Conclusions

We have introduced an approach to Open IE based on syntactic constraints over POS tag sequences targeted at Spanish language. We described the rules for relation phrases and their arguments in Spanish and their implementation in EXTRHECH system. Further, we presented a series of experiments with EXTRHECH and showed (1) that the performance of this approach to Open IE is similar for English and Spanish, and (2) that EXTRHECH's performance is robust on texts of varying quality. We also gave a brief classification of errors by their types and causes.

Our future plans include implementation of shallow parsing and syntactic n-grams (Sidorov et al., 2012; Sidorov et al., 2013; Sidorov et al., 2014; Sidorov, 2013a; Sidorov, 2013b), as well as learning techniques, and analysis of their influence on the system's performance.

Acknowledgments

The work was partially supported by the Government of Mexico: SIP-IPN 20144534 and 20144274, PIFI-IPN, and SNI. We thank Yahoo! for travel and conference support for this paper.

References

Honorato Aguilar-Galicia. 2012. Extracción automática de información semántica basada en estructuras sintácticas. Master's thesis, Center for Computing Research, Instituto Politécnico Nacional, Mexico City, D.F., Mexico.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36. Association for Computational Linguistics, June.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, pages 2670–2676.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Joseph Clancy Clements. 2006. Primary and secondary object marking in Spanish. In J. Clancy Clements and Jiyoung Yoon, editors, *Functional approaches to Spanish syntax: Lexical semantics, discourse, and trasitivity*, pages 115–133. London: Palgrave MacMillan.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December.

Oren Etzioni. 2011. Search Needs a Shake-Up. *Nature*, 476(7358):25–26, August.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, ROBUS-UNSUP '12, pages 10–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher Horn, Alisa Zhila, Alexander Gelbukh, Roman Kern, and Elisabeth Lex. 2013. Using factual density to measure informativeness of web documents. In *Proceedings of the 19th Nordic Conference on Computational Linguistics*, NoDaLiDa.

Marshall Kirkpatrick. 2011. New 5 billion page web index with page rank now available for free from common crawl foundation. `http://readwrite.com/2011/11/07/common_crawl_foundation_announces_5_billion_page_w`, November. [last visited on 25/01/2013].

Geoffrey Leech and Andrew Wilson. 1999. Standards for tagsets. In *Syntactic Wordclass Tagging*, pages 55–80. Springer Netherlands.

Thomas Lin, Mausam, and Oren Etzioni. 2010. Identifying functional relations in web text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1276. Association for Computational Linguistics, October.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *EMNLP-CoNLL*, pages 523–534. ACL.

Lluís Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India, February. Global Wordnet Conference 2010, Narosa Publishing House.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2012. Syntactic dependency-based n-grams as classification features. In M. González-Mendoza and I. Batyrshin, editors, *Advances in Computational Intelligence. Proceedings of MICAI 2012*, volume 7630 of *Lecture Notes in Artificial Intelligence*, pages 1–11. Springer.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2013. Syntactic dependency-based n-grams: More evidence of usefulness in classification. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2013*, volume 7816 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

Grigori Sidorov. 2013a. Non-continuous syntactic n-grams. *Polibits*, 48:67–75.

Grigori Sidorov. 2013b. Syntactic dependency based n-grams in rule based automatic english as second language grammar correction. *International Journal of Computational Linguistics and Applications*, 4(2):169–188.

Stephen Soderland, Brendan Roof, Bo Qin, Shi Xu, Mausam, and Oren Etzioni. 2010. Adapting open information extraction to domain-specific relations. *AI Magazine*, 31(3):93–102.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alisa Zhila and Alexander Gelbukh. 2014. Automatic identification of facts in real internet texts in Spanish using lightweight syntactic constraints: Problems, their causes, and ways for improvement. *Submitted*.