ISSN 1870-4069

An Approach for Knowledge Graph Construction from Spanish Texts

Andrea G. Garcia Perez[1], Ana B. Rios Alvarado[1],
Tania Y. Guerrero Melendez[1], Edgar Tello Leal[1], Jose L. Martinez Rodriguez[2]

[1] Autonomous University of Tamaulipas,
Mexico

[2] CINVESTAV Unidad Tamaulipas,
Mexico

agidaltig@gmail.com, {arios, tyguerre, etello}@docentes.uat.edu.mx,
lmartinez@tamps.cinvestav.mx

Abstract. Knowledge Graphs (KGs) represent valuable data sources for the Educational and Learning field, allowing computers and people to process and interpret information in an easier way. In addition KGs are useful to get information by software systems on any data resource (person, place, organization, among others). KGs are represented through triples composed of entities and semantic relations commonly obtained from textual resources. However, exploiting triples from text is complicated due to linguistic variations, where Spanish has been slightly approached. This paper presents a method for constructing KGs from textual resources in Spanish. The method is composed of four stages: 1) obtaining of documents, 2) identification of entities and their relations, 3) association of entities to linked data resources, and 4) the construction of a schema for the KGs. The experiments were performed over a set of documents in a general and computer science domain. Our method showed encouraging results for the precision in the identification of entities and semantic relationships for the KGs constructed from unstructured texts.

Keywords: knowledge graph, construction of graph from texts.

1 Introduction

In recent years the Web has become a global repository that represents a source of knowledge in a wide range of domains and languages, where the information is shared and stored in different formats. The whole range of data is attractive for different commercial, industrial, and academic institutions. However, extracting and processing information coming from sources such as the Web is not a straightforward task even if it is manually done by a human. There is a growing desire for direct and automatic access to data. For example, in the educational domain, where the data need to be modeled by a representation that facilitates the comprehension, sharing, and reusing of information by users (and systems).

pp. 9–17; rec. 2019-09-01; acc. 2019-09-24

One way to represent a data model is through the Linked Data repositories, which collect resources about different topics that can be interconnected and satisfy diverse standards of publication and exchange of data.

The use of these standards promotes to share and query information, allowing the identification and extraction of new knowledge through the entities (objects of the real world) and the relationships between them [4], which produces *Knowledge Graphs* (KG).

KGs have been adopted as data representation models in applications such as DBpedia, YAGO, Wikidata, among others. A KG can be constructed to store data in a triple store that can be analyzed to acquire or infer new knowledge about entities (e.g., people, places, organizations) and concepts, as well as interconnected entities.

The representation of text as KGs is very important in education or learning environments because KGs allow people to achieve better understanding, interoperability, and accessibility of data [2, 3]. Thus, for exploiting large amounts of knowledge, some approaches obtain a graph model to represent entities and semantic relations [8, 9, 5, 2, 1]. Most of them use NLP techniques for extracting relations in English language, and in some cases belong to an specific domain such as Education [2, 1], Medicine [9, 8], and Security [7]. However, extracting triples (and KGs) from text is a complicated task, where the text may require patterns and rules to obtain entities and semantic relations. Additionally, even though there are over 583 million Spanish speakers[1] in the world and the production of text (e.g.,by collaborative networks, blogs, social networks, emails, and other applications) is increasing daily, this language has been scarcely explored for the extraction and representation of KGs.

In this paper, we propose a method to automatically build KGs from unstructured texts in Spanish. Our method is composed of four stages: 1) obtaining and processing the documents, 2) identification of entities and their relations by lexical patterns using NLP tools, 3) association of entities to linked data resources, and 4) the construction of a schema for the KGs. The entities and relationships are extracted from text in a process which can be performed by lexical patterns and transformed to an RDF representation.

This paper is organized as follows. Section 2 describes the proposed method, starting with the obtaining and processing of text documents in Spanish and then with the identification of entities and semantic relationships (incorporating a method to associate them with DBpedia resources). Next, in section 3 the experiments and results are shown. Finally, Section 4 presents the conclusions and further work.

2 Computational Methodology

Our proposed method aims to generate a document model representation in the form of a KG from a set of general domain unstructured documents written

[1] https://www.ethnologue.com/language/spa

in Spanish. The method is composed of four main steps described in the next subsections.

2.1 Step 1: Obtaining and Processing the Documents

In the first step, each document contained in a collection is prepared for the extraction of entities and semantic relations. For this, the following tasks are needed:

- *Cleaning the document.* The document is processed as UTF-8 format, dealing with errors of accents and special characters.
- *Splitting the text in sentences.* The text is divided into sentences since it is the grammatical structure that allows extracting more relationships with higher precision than other textual units.
- *Grammatical tagging.* Using a grammatical parser, a label (e.g., noun, verb, article) for each word in the sentence is obtained. The labels provided by this process allow the generation of lexical patterns that will identify the semantic relationships between nouns, verbs, determinants, and other grammatical components.
- *Named-entity identification.* We use a Named Entity Recognition parser to identify named entities such as places, people, and organizations.

2.2 Step 2: Identification of Entities and Semantic Relationships

The second step requires the tags obtained by the grammatical tagging and named-entity recognition for extracting entities and semantic relationships (taxonomic, non-taxonomic and structural relations). Thus, it is possible to extract the elements described below:

- *Entities.* The words that represent simple nouns, compound nouns, named entities, and specialized terms are considered as entities, which may contain adjectives, determinants and/or prepositions. We considered a set of lexical patterns to recognize entities with more than one word. For example, the sequence *Magdalena Carmen Frida Kahlo Calderon* is identified.
- *Taxonomic relations.* It classifies a specific concept as part of a more general concept or indicate if an entity has a type. For example, <*FridaKahlo, rdf:type, Artista/Artist*>, where *Frida Kahlo* is a type of *Artist* class.
- *Equivalence relations.* These relationships establish the equality or equivalence between two expressions that are apparently different. For example, *infancia/childhood*[2] is synonym of *niñez/childhood* (in English language both concepts are related with *childhood*).
- *Structural relations.* These relationships describe how a concept (or set of concepts) can be broken down into parts of subsystems. For example,<*puerta de embarque/gate, isPartOf, Aeropuerto/Airport*>, where *Airport* is composed of *gates* and other elements.

[2] Word in Spanish/Word in English

- *Non-taxonomic relationships.* It is a kind of semantic relationship in which two or more entities are linked in a non-hierarchical structure. It can represent an action, event in time or location in space. For example, <*FridaKahlo, nacidaEn/wasBornIn, Coyoacan*> is used to indicate that a person named *Frida Kahlo* was born in a place called *Coyoacan.*

Additionally, a set of lexical patterns and regular expressions are required for identifying entities like *tópicos de creciente interés/topics of growing interest* or structural relations as *el cuerpo humano esta compuesto de células/the human body is composed by cells.*

2.3 Step 3: Association of Entities to a Linked Data Repository

The representation of KGs relies on the Resource Description Framework (RDF[3]). In this data model both the resources being described and the values describing them are nodes in a directed labeled graph. The arcs connecting pairs of nodes correspond to the names of the property types. Thus, a semantic triple is a set of three elements that codifies a statement about semantic data in the form of subject–predicate–object expressions.

We obtain RDF triples using the entities obtained in step 2 and querying property data from a linked data repository, in our case DBpedia. We consider some relevant DBpedia properties: *rdf:type*, *dcterms:subject*, *rdf:seeAlso*, and *owl:sameAs.* These properties are retrieved for all the identified entities because they represent taxonomy, categorization (or type), and equivalence relations. Restricting the association or linked entities with the mentioned properties is a controlled form of enriching the graph without irrelevant or ambiguous information.

We constructed SPARQL[4] queries for consulting the properties in DBpedia. In Listing 1.1 is represented a query associated to the relation *dcterms:subject* for retrieving the resources (*?uri*) related with a specific entity (*?val*):

- SELECT
- ?uri
- ?val
- WHERE ?uri dcterms:subject ?val .

2.4 Step 4: Knowledge Graph Construction

The construction of the KG is divided into two tasks:

- *Scheme specification.* When semantic relations are obtained from linked data repositories as DBpedia, vocabularies with already defined properties are used. In the case of triples extracted from the text, the identified properties or predicates must be specified. For this reason, we define a scheme for such properties. However, only the label of the property is defined because it is not known the range of objects or resources that will use this property.

[3] https://www.w3.org/RDF/

[4] https://www.w3.org/TR/sparql11-overview/

- *Triples construction.* Once the scheme of is specified, it is possible to build the KG using an RDF format, in which the resources and the identified relationships are described (either textual relations or retrieved using DBpedia).

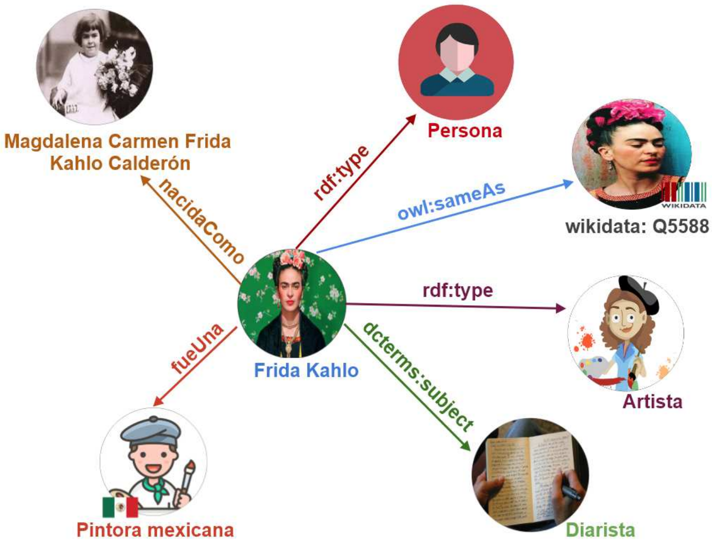

Fig. 1. Example of KG for the entity "Frida Kahlo".

The generated RDF document corresponds to the representation of the processed document, so the KG represents the context and knowledge of the topics described in a set of documents. For each document, an RDF document is constructed.

An example of KG is shown in Figure 1, where the *FridaKhalo* entity is related by using the properties *rdf:type*, *owl:sameAs*, *dcterms:subject* associated with DBpedia, and *wasA*, *named* (in Spanish, *fueUna*, *nacidaComo*, respectively) are data properties extracted from text:

1. <FridaKahlo, *rdf:type*, Persona>
2. <FridaKahlo, *owl:sameAs*, wikidata:Q5588>
3. <FridaKahlo, *rdf:type*, Artista>
4. <FridaKahlo, *dcterms:subject*, Diarista>
5. <FridaKahlo, *fueUna/wasA*, Pintora Mexicana>
6. <FridaKahlo, *nacidaComo/named*, Magdalena Carmen Frida Kahlo Calderon>

3 Experiments and Results

One way to evaluate the KG is to determine the quality of the extracted semantic relationships and entities used to build the graph. However, validating the results of the extraction of such elements can be a complex task due to the limitation of published datasets for the construction of KG in Spanish.

The extraction of the elements can be seen as an information retrieval task, so it can be evaluated with the measures Precision, Recall, and F-measure [6]. Precision evaluates the capability of the method to exclude those elements that do not represent true entities or semantic relationships and Recall assesses the capability to recover all those entities or semantic relationships among which we know that exists. For its part, F-measure is a harmonic average between Precision and Recall and it reflects the performance of the method.

The experiments were conducted to evaluate two stages: 1) identification of entities and semantic relations from text (according to lexical patterns), and 2) identification of each one of the triple elements (subject, predicate, and object); in the case of subject and object their association with a resource in DBpedia for KG construction.

3.1 Dataset

We constructed the dataset used in the experiments, it contains nineteen documents, nine about a general domain and ten about computer science domain, with a total of 250 sentences and 127 semantic relations, in the Spanish language. Each document has an identifier corresponding to a number from 1 to 19. Three computer science domain experts manually identified a set of entities and triples (subject, property, and object) from sentences. We considered entities of class place, people and those resources that are in DBpedia[5] in Spanish according to a human expert.

3.2 Identification of Entities and Semantic Relationships

For the entities identification process individually or into the semantic relationships the following features were considered: extracting all the entities that represent a specialized term, a named entity, simple singular nouns, or compound nouns and matching the sentences with lexical patterns for extracting taxonomic, equivalent, structural, and non-taxonomic relationships. Experiment 1 involves taking account that each sentence match with the lexical patterns. Experiment 2 involves to identify each component of the triple (subject, predicate, and object) given a sentence.

To evaluate the entity extraction process, the domain experts annotated 137 entities for a document. The results of Precision, Recall and F-measure are shown in Table 1. Using the lexical patterns it is possible to obtain entities like *la población del norte de pingüinos Rockhopper/the northern population of Rockhopper penguins*, and *especies de pingüinos/species of penguins.* Such detected entities do not have an associated resource to DBpedia but they represent important concepts into text and they have some semantic relation. On the other hand, the entities *biodiversidad* and *hemisferio sur*, also detected by lexical patterns,

[5] http://es.dbpedia.org/

were associated with the resources *http://es.dbpedia.org/page/Biodiversidad* and *http://es.dbpedia.org/page/Hemisferio_sur*, respectively. As shown in Table 1, when fewer patterns are considered, the number of incorrect identified decreased. However, less than 25% of real relationships is obtained. When more patterns are considered, the number of incorrect increases but precision and recall values are balanced.

Table 1. Entity extraction results.

	Experiment 1	**Experiment 2**
Entities	137	137
Precision	1.00	0.91
Recall	0.25	0.62
F-measure	0.40	0.74

The results of the relationship extraction process are described in Table 2, where a comparison between the results of both experiments is presented. Given the number and kind of patterns considered for Experiment 1, there are no results for all the documents in that test, which means that in documents 2,3,4,5,6, and 7, entities were not obtained.

In contrast, in documents 1,8, and 9 the entities identified are all correct, getting a Precision of 1.0 but a low Recall and also a low F-measure. Nevertheless, in Experiment 2, all documents obtained results for extracting entities and relations tasks, but only one document had a precision of 1, this was because the number of matches was bigger than other documents.

Table 2. Summary of results for the relationships extraction process.

	Experiment 1	**Experiment 2**
Total correct	5	49
Total incorrect	0	23
Total identified	5	75
Total real	84	84
Average precision	1.00	0.65
Average recall	0.06	0.58
Average F-measure	0.31	0.82
Median precision	1.00	0.67
Median recall	0.18	0.60
Median F-measure	0.31	0.60

For the extraction of semantic relations as triples, it was evaluated the Precision, Recall and F-measure in the identification of subject, predicate, and object. A subject or an object can be associated with a resource in DBpedia.

Three human experts manually identified the correct subjects and objects related to the context of the documents.

In general, the results for the method are related to the patterns considered in the identification of entities and the style of writing in the texts. There could be texts in which the author describes some concepts and the method will be benefited. However, the method was designed for general purpose or general domain so it cannot be benefited for any particular writing style.

4 Conclusions and Further Work

In this paper, a method for the construction of knowledge graphs was presented. This method consists of four phases, the last of them focused on the construction of the model.

One of the most important phases is the stage of extraction of entities and semantic relationships and the alternatives to perform it out, the lexical patterns and the results by applying each one of the alternatives. For the entity extraction, it is considered the occurrence of nouns accompanied by other elements that end in simple entities, multiple entities or with a sequence of nouns and concepts, also considering entities that represent specialized terms, among others.

With respect to the semantic relationships, the proposed method is capable of identifying and extracting non-taxonomic, taxonomic, equivalence and structural relations, which represent triples in the knowledge graph. Our method showed encouraging results for the precision in the identification of entities and semantic relationships.

As future work, we plan to extend the experiments to consider the tasks of entity and semantic relation extraction from text and the further step of linking such elements with resources from a knowledge base. Moreover, we also plan to apply the extracted KGs in tasks such as information retrieval and data visualization to measure the degree to which linked data and KGs can support the Educational and Learning field.

Acknowledgements. This work was partially supported by the National Council of Science and Technology (CONACYT) and Basic Science Research Project SEP-CONACYT of Mexico under Grant 256922.

References

1. Buscaldi, D., Dessı, D., Motta, E., Osborne, F., Recupero, D.R.: Mining scholarly data for fine-grained knowledge graph construction. In: Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) at ESWC2019 (2019)
2. Chen, P., Lu, Y., Zheng, V.W., Chen, X., Yang, B.: Knowedu: A system to construct knowledge graph for education. IEEE Access 6, 31553–31563 (2018), https://doi.org/10.1109/ACCESS.2018.2839607
3. Cui, J., Yu, S.: Fostering deeper learning in a flipped classroom: Effects of knowledge graphs versus concept maps. BJET 50(5), 2308–2328 (2019), https://doi.org/10.1111/bjet.12841

4. Hernández-Pérez, T.: En la era de la web de los datos: primero datos abiertos, después datos masivos. El profesional de la información (EPI) 25(4), 517–525 (2016)
5. Hu, S., Zou, L., Yu, J.X., Wang, H., Zhao, D.: Answering natural language questions by subgraph matching over knowledge graphs. IEEE Transactions on Knowledge and Data Engineering 30(5), 824–837 (2018)
6. Serra, I., Girardi, R.: A process for extracting non-taxonomic relationships of ontologies from text. vol. 3, p. 119. Scientific Research Publishing (2011)
7. Szekely, P., Knoblock, C.A., Slepicka, J., Philpot, A., Singh, A., Yin, C., Kapoor, D., Natarajan, P., Marcu, D., Knight, K., et al.: Building and using a knowledge graph to combat human trafficking. In: International Semantic Web Conference. pp. 205–221. Springer (2015)
8. Yu, T., Li, J., Yu, Q., Tian, Y., Shun, X., Xu, L., Zhu, L., Gao, H.: Knowledge graph for tcm health preservation: Design, construction, and applications. Artificial intelligence in medicine 77, 48–52 (2017)
9. Zaki, N., Tennakoon, C., Al Ashwal, H.: Knowledge graph construction and search for biological databases. In: 2017 International Conference on Research and Innovation in Information Systems (ICRIIS). pp. 1–6. IEEE (2017)