Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 43-54 recibido 08-11-2024 revisado 20-01-2025 aceptado 07-02-2025

Tailoring a Knowledge Discovery Framework to Process Pharmacologic Documents

Adaptación de un Marco de Descubrimiento de Conocimiento para Procesar Documentos Farmacológicos

Isabel Moreno,[1] **Alejandro Piad-Morffis,**[2] **Yoan Gutiérrez,**[1] **Paloma Moreda**[1]
[1]Universisy of Alicante, Spain
[2]University of Havana, Cuba
{imoreno, ygutierrez, moreda}@dlsi.ua.es, apiad@matcom.uh.cu

Abstract: This paper introduces a specialized knowledge discovery framework designed to process health technical documents and extract knowledge. The framework improves existing technology, known as LETO, through the integration of CARMEN, a multilingual entity classification system capable of infusing health-related semantics into the initial versatile approach. This collaborative approach enables the generation of domain-specific knowledge graphs for two languages, Spanish and English. Additionally, this provides a valuable means by which to explore relationships within the health domain that could otherwise remain undiscovered. The resulting technology is subjected to an evaluation procedure using standard metrics employed in knowledge discovery tasks, illustrating how CARMEN contributes to an augmentation in the knowledge discovered in LETO. Thus, the generated knowledge graph can be leveraged for the creation of explanatory representation techniques, facilitating a more comprehensive articulation of human knowledge and potentially serving, among other purposes, as an educational resource.
Keywords: Knowledge Discovery, Health Documents, Multilingual Entity Classification, Knowledge Augmentation.

Resumen: Este trabajo presenta un marco especializado de descubrimiento de conocimiento diseñado para procesar documentos técnicos de salud. Este mejora la tecnología existente, conocida como LETO, mediante la integración de CARMEN, un sistema de clasificación de entidades multilingüe capaz de incorporar semántica relacionada con la salud. Este enfoque colaborativo permite la generación de grafos de conocimiento específicos del dominio en dos idiomas, español e inglés. Además, ofrece un medio valioso para explorar relaciones dentro del ámbito de la salud que, de otro modo, podrían permanecer sin descubrir. La tecnología resultante se somete a un procedimiento de evaluación utilizando métricas estándar empleadas en tareas de descubrimiento de conocimiento, demostrando cómo CARMEN contribuye a aumentar el conocimiento descubierto en LETO. Así, el grafo de conocimiento generado puede aprovecharse para usos de representación explicativa, facilitando una articulación más completa del conocimiento humano y, entre otros fines, sirviendo potencialmente como un recurso educativo.
Palabras clave: Descubrimiento de Conocimiento, Documentos Técnicos de Salud, Clasificación de Entidades Multilingüe, Integración Semántica.

ISSN 1135-5948 DOI 10.26342/2025-74-3 ©2025 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Knowledge Discovery (KD) analyzes large data corpora to automatically extract or synthesize useful knowledge (Fayyad et al., 1996). Among various data sources, natural language text is particularly challenging due to its unstructured and subjective nature. In the medical domain, complex technical terms add further difficulties (Roberts, 2017).

KD approaches fall into two paradigms: information extraction (IE) and ontology learning (OL) (see section 2). IE extracts explicit facts from text, while OL infers abstract concepts and relationships. Combining both enhances knowledge extraction quality.

This research integrates a general-purpose KD framework with an IE approach for the health domain. Specifically, we select LETO (Estevez-Velarde et al., 2019), a general KD framework, and CARMEN (Moreno, Romá-Ferri, and Moreda, 2019), a language- and domain-independent entity classification system, both easily adaptable to new scenarios.

LETO recognizes relevant information in natural language text via software components called "sensors", which is later refined, aggregated, and abstracted to construct knowledge graphs that represent the most relevant concepts and relations in a given domain. However, LETO cannot leverage domain-specific knowledge, such as custom entity or relation labels. For this reason, the knowledge extracted with LETO alone is too abstract and lacks details relevant to domain experts in highly sensitive domains such as health.

In contrast, CARMEN is a supervised machine learning system for named entity recognition that does not rely on domain-specific rules or external knowledge. It just needs to be trained in the desired languages and can naturally be extended to any domain with custom entity and relation labels, as long as a suitable corpora is available.

As a case of study, this research focuses on the health domain. Specifically, we target standardized and official medical documentation comprised of product summaries, public assessment reports, and other types of technical documents that describe medical conditions, products, and procedures. The main objective of this research is thus to demonstrate how the combination of LETO and CARMEN enables the discovery of relevant medical information in the aforementioned domain.

To tackle this problem, we propose to extend LETO with a new sensor, based on the CARMEN architecture, focused on entity classification of pharmacologic content for two widely spoken languages (Vítores, 2017), i.e. Spanish and English. CARMEN is selected to fit in the proposed LETO pipeline because it is able to deal with both languages and to learn health-relevant knowledge, namely: (i) disorders, (ii) chemicals and drugs, as well as (iii) procedures.

The hypothesis of this research is that adding an extra entity tagger (i.e. CARMEN) as a new sensor to deal with domain-specific entities will increase the KD recall, while maintaining LETO's OL architecture.

The key contributions of this work can be summarized as follows:

- The development of a novel knowledge discovery pipeline integrating a general-purpose framework (LETO) with a domain-specific entity classification system (CARMEN).
- A demonstration of how domain-specific entity tagging enhances the knowledge discovery recall of LETO in the health domain.
- The creation and publication of knowledge graphs from health documents that can be leveraged for explanatory representation techniques and educational resources.

The remainder of this paper is structured as follows: Section 2 provides a review of related work in knowledge discovery, information extraction, and ontology learning. Section 3 details our proposed framework, including the architecture of LETO, the integration of CARMEN, and the entity classification and filtering processes. Section 4 describes the experimental setup, datasets, and evaluation metrics used to assess the performance of our framework. Section 5 presents and discusses the experimental results, highlighting the impact of CARMEN on knowledge discovery recall. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Literature Review

Discovering useful knowledge in text involves several tasks that range from the identification of syntactic structures that indicate relevant fragments of knowledge (e.g., recognizing

mentions of relevant entities in a sentence) to the interrelation of semantically connected elements (e.g., normalizing variants of entity names) and furthermore to the application of inference rules to obtain new knowledge. In this wide range of tasks, two general approaches are commonly distinguished: information extraction (IE) and ontology learning (OL) (Feldman and Sanger, 2007; Asim et al., 2018).

IE approaches often deal directly with textual sources and attempt to extract only explicit information directly mentioned in the text(Feldman and Sanger, 2007). Health-related approaches analyze scientific articles, patient reports or technical documents aiming at recognizing and classifying medical entities(Boag et al., 2018) (e.g. 'headache' is a disease), linking these entities to health knowledge bases (Savova et al., 2010; Friedman et al., 2004), such as UMLS (e.g. 'headache' is represented in UMLS with CUI 'C0018681'), or extracting attributes from medical entities (Viani et al., 2018) (e.g. Heart rate from an Electrocardiogram).

OL approaches handle the knowledge end of the spectrum, i.e., the construction of formal knowledge representations that can be used for automatic inference, often stored in a knowledge base or ontology(Asim et al., 2018). These techniques can automatically discover knowledge from unstructured content and create ontologies as a result of the processing task, such that the final refinement step consists of inferring classes, instances, and semantic relations.

Since IE and OL aim to find knowledge, both KD approaches can be seen as complementary. IE converts the unstructured textual information into a problem-specific representation, such as a list of relevant entities. Next, the OL builds a general-purpose semantic representation, such as an ontology. Although semantic representations of OL provide more meaningful information(Konys, 2018; Estevanell-Valladares et al., 2021), IE is the focus of most KD techniques applied in the health domain(Patrick et al., 2011). This may be due to the high dependency of OL techniques on the target domain(Wong, Liu, and Bennamoun, 2012; Konys, 2018) (e.g. making heavy use of domain-specific external knowledge or ad hoc patterns), thus lacking the necessary strategies to be extended to different domains easily.

3 Proposal

This work specializes an existing KD framework to extract health-related information by extending general knowledge triplets with domain knowledge. It builds on a previous KD pipeline (Galiano et al., 2023; Estevez-Velarde et al., 2018), enhancing it with domain-specific entity classification and filtering for the health domain.

3.1 LETO: General-Purpose Knowledge Discovery Architecture

Figure 1 shows a schematic representation of our extended pipeline. The overall KD pipeline proposed by LETO consists of three phases, namely: Sensorial, Structural and Knowledge. This pipeline extracts entities and relations (in the form of Subject-Action-Target triplets) from text, and progressively organises and refines this information ending with the construction of an ontology with the most relevant elements. Since this framework is designed for general-purpose KD, it fails to discover relevant entities in highly specialized domains such as health. For this reason, our proposal enriches this existing general-purpose pipeline (blue boxes in Figure 1) with the addition of a domain-specific plugin (i.e. CARMEN - lilac boxes in Figure 1) that consists of two components: an entity classifier trained on health corpora, and a filtering component, specifically designed for this research, to remove unwanted domain entities.

The resulting process consists of the following phases, mixing general-purpose components and domain-specific CARMEN components:

Sensorial Level: In the first phase, Sensorial Level in Figure 1, Subject-Action-Target triplets are extracted from each sentence. This process is based on heuristic rules that consider grammatical and syntactic features of the text. At this point, Subject and Target entities are assigned one of several standard named entity recognizer labels (e.g. person, organization, date, etc.) using the general-purpose entity tagger already available in LETO, namely spaCy (Explosion, 2019).

Since spaCy only recognizes a small subset of entity types (e.g., person, organization, location, etc.), LETO relies not only

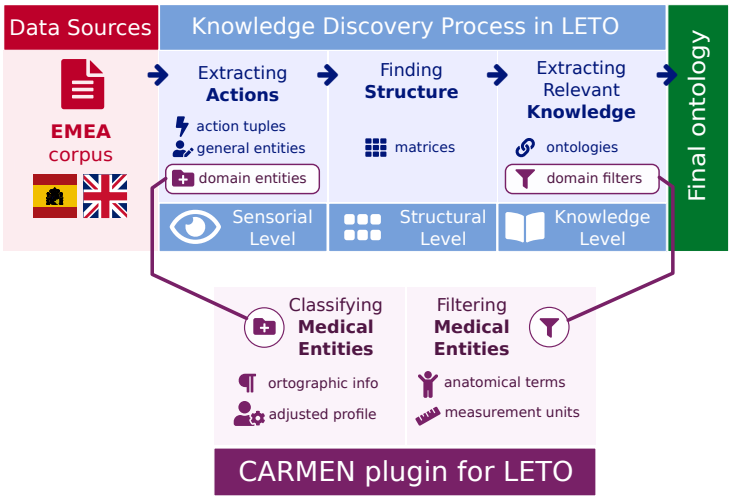

Figure 1: Schematic representation of our extended knowledge discovery process. Input is colored in red, previous pipeline in blue (see (Estevez-Velarde et al., 2018, Fig. 1)), medical extension in lilac, and result in green.

on spaCy entity labels, but also considers dependency-based features to allow capturing triplets that do not contain labelled spaCy entities. This is achieved by applying a set of pattern rules on the structure of the dependency tree, e.g., considering noun-verb-noun and similar structures.

We propose enriching this phase with a domain-specific entity classifier from CARMEN. For this purpose, entities that are not recognized as one of the standard types in LETO are further processed by CARMEN to assign a domain-specific label. This step is described in more detail in the next section.

Structural Level: Structural Level in Figure 1, the concepts (i.e., Subjects, Actions and Targets) are combined in a common representation that encodes their interrelations. This representation consists of three matrices, one for each pair of concept types (e.g., Subjects and Actions, Subjects and Targets, etc.). Each matrix relates pairs of concepts that appear together in at least one action triplet.

Knowledge Level: Finally, in the third phase, Knowledge Level in Figure 1, the most relevant knowledge is selected to create an ontology automatically. This process is performed in LETO using a clustering algorithm (i.e., Affinity Propagation (Frey and Dueck, 2007)) for compacting purposes. This assumes that words sharing similar relations can be merged.

We propose enriching this process with a filtering component applied to filter subjects and targets that represent noise, since we know that they belong to undesired semantic types beforehand. This process is described in more detail in the next section 3.2.

For more details about LETO, we refer the reader to the original article(Estevez-Velarde et al., 2019).

3.2 CARMEN: Plugin for Entity Classification and Filtering

As previously mentioned, CARMEN is the system taken to classify health-specific concepts. This system is language and domain independent, and was tailored for this particular scenario. In this manner, a general purpose KD system (i.e. LETO) is enhanced to deal with concepts specific to this health-related domain. To achieve this goal there are two phases: (i) classify medical entity mentions in the Sensorial Level; and, (ii) filter medical entity mentions of measurements and anatomy related concepts in the Knowledge Level. Next, both phases are further described:

1. **Classifying Medical Entities:** CARMEN is able to automatically determine the optimal set of features that represent entity mentions, as well as the associated supervised machine learning algorithm,

for a given training corpus. In this case, it was adjusted for Spanish to obtain four optimal configurations (i.e. set of parameter's values) to be employed for any language depending on constraints and available resources, namely: (i) CARMEN$_O$ that only takes into account orthographic information; (ii) CARMEN$_{PR}$ that takes into account orthographic and contextual information; (iii) CARMEN$_K$ that takes into account orthography and contextual information, as well as external knowledge; and (iv) CARMEN$_{PO}$ that integrates a post-process to CARMEN$_K$.

Ortographic information considers features such as lemma, affixes, token length, etc. Contextual information enriches the features of a token with features from nearby tokens within a fixed window. External knowledge is based on existing lexicons and taxonomies of medical terms.

Aiming at reducing bias and having equal conditions, external knowledge configurations (i.e. CARMEN$_K$ and CARMEN$_{PO}$) are not taken into account. This decision is motivated by the fact that not all terminologies are available for all languages and the coverage is not equal for all languages. Among the two remaining options, the one with higher results, CARMEN$_{PR}$, was selected to be used for Spanish and English in order to have equal conditions, thereby providing a fair comparison among languages.

New texts, in which entity mentions have been previously identified by the Sensorial level (see Figure 1), apply the CARMEN entity classification process. It consists of two stages:

Entity Representation: The aim is to convert each mentioned entity into a set of features (F). The feature set includes local information of mention, inspired by state-of-the-art, namely: mention, character n-grams and affixes (i.e. suffixes and prefixes) of the entity. As a novelty, this feature set includes the context of mentions through an ensemble of profiles (Lopes and Vieira, 2015). This ensemble consists of three profiles, one per entity type, namely: (i) disorders, (ii) chemicals and drugs, as well as (iii) procedures. Each profile P_{type} can be seen as a distributed vector composed of a set of pairs: (i) one descriptor d, representing the name of the component; and (ii) its relevance idx, a numerical value calculated using the *term frequency, disjoint corpora frequency* (Lopes and Vieira, 2015) - TFDCF - index. Context information considers as features both name and TFDCF.

Machine Learning: In this step every entity mention (i.e. the feature set F previously generated) receives one of the types from all the three possible ones. It applies the random forest (Breiman, 2001) algorithm with 60 trees.

The overall functioning of this phase is described in Figure 2.

2. **Filtering Medical Entities:** For the entity filtering step in phase 3 (Knowledge Level in Figure 1), a domain-specific dictionary of entities is built from external terminologies and used to remove unwanted entities. Specifically, it removes concepts belonging to the anatomical semantic group (i.e. body parts, location and systems) from UMLS semantic network (McCray, Burgun, and Bodenreider, 2001) and units of measurement. Body related terms are gathered from the Mantra terminology (Kors et al., 2015), a customized subset of UMLS (Bodenreider, 2004) Metathesaurus (2017AA-full version) that comprises SnomedCT (IHTSDO, 2017) as well as MeDRA (Brown, Wood, and Wood, 1999) and MeSH (Rogers, 1963) terminologies. The units of measurement dictionary was compiled from SNOMED-CT as in(Moreno et al., 2017).

4 *Experimental Evaluation*

In order to tackle the evaluation, we focus on four criteria to gauge the ontology learning process:

Coherence: Evaluation of the triplets obtained based upon their consistency with the domain of the corpus. For this purpose we randomly select 100 triplets, following (Estevez-Velarde et al., 2018). This subset is manually evaluated by 3

Algorithm 1 CARMEN entity classification

```
—— Parameters for the adjustment
G_S ← Gold standard to optimize                          ▷ MANTRA_SPANISH
R ← Restrictions          ▷ PR: Orthographic information and adjusted profiles
—— Parameters for the training
G_EN ← English Gold standard                             ▷ MANTRA_ENGLISH
G_SP ← Spanish Gold standard                             ▷ MANTRA_SPANISH
—— Parameters for the classification of new texts
Text_EN ← English input text with entity mentions identified
Text_SP ← Spanish input text with entity mentions identified

—— Step 1: Adjustment
C ← Find the best configuration using G_S, R                          ▷ CARMEN_PR

—— Step 2: Training
M ← []                                                          ▷ Trained models
for g ∈ G_EN, G_SP do
    F ← []                                  ▷ Feature set matrix of the training data
    T ← []                                          ▷ Labels of the training data
    for e ∈ g do                          ▷ For each mention in the training data
        F_e ← GETREPRESENTATION(e, g, C)                  ▷ Entity representation
        T_e ← e.GETLABEL()                          ▷ Label in the training data
    M_g ← train(F, T)                        ▷ Train the model in each language

—— Step 3: Classification of new texts in both languages
lang ← EN, SP
F ← []                                      ▷ Feature set matrix of the new data
T ← []                                    ▷ Predicted labels of the the new data
for L ∈ lang do                                              ▷ For each language
    t ← Text_L
    for e ∈ t do                                  ▷ For each entity mention identified
        F_e ← GETREPRESENTATION(e, t, C)                  ▷ Entity representation
        T_e ← M_L.CLASSIFY(F_k)                                  ▷ Machine learning

return T                                  ▷ Predicted labels of the the new data
```

Figure 2: Classifying Medical Entities with the CARMEN Algorithm.

human raters to assign a score between 1 and 3. The value 1 is considered disagreement between the original corpus and the resulting triplet, 2 agreed and 3 strongly agreed. The average score assigned to each triplet is computed, and the percentage of triplets in each category is reported.

Coherence without Domain Knowledge: The previous manual evaluation is carried out using the original knowledge discovery pipeline (Estevez-Velarde et al., 2018), without the CARMEN-based sensor, so as to quantify the degree of improvement made by our contribution. The same process as before is applied, and the resulting averages are compared with the results from the extended pipeline (i.e., LETO+CARMEN). The absolute difference between each category (i.e., disagree, agree, strongly agree) is reported (see Table.

Reduction Rate: Analysis of the reduction of knowledge produced by the clustering algorithm. For this purpose, we analyse the distribution of the number of concepts and relations before and after the Knowledge Level phase to detect the most significant sources of reduction. A histogram of the average number of triplets in which each entity participates is computed, before and after the clustering process. The expected result is that after the clustering process, a smaller number of total entities remain in the ontology, each of which appears in a larger number of triplets, thus indicating an aggregation of semantically similar concepts (see Figure 3).

Knowledge Domain Growth: Comparison of the relevance of the general-purpose and the domain-specific knowledge extracted. To this end, we measure the relative number of domain-specific concepts with respect to the overall size of the ontology. A comparison between the number of standard entities and

actions recognized by LETO and the novel domain entities recognized by the CARMEN-based sensor is reported, along with percent values.

Comparison across Languages: The preceding analyses are scrutinized for the English and Spanish languages to determine if there are any variations in performance or quality of the extracted knowledge based on the input language.

4.1 Corpora

Two types of data were used in our experiments:

Annotated Corpus: CARMEN was trained on the Mantra gold standard (Kors et al., 2015). This annotated corpora comprises five languages, but this work only gathers parallel documents available in Spanish and English (see Table 4.1). It is composed of 86 drug labels coming from the European Medicines Agency (EMEA); and 100 MedLine titles per language. It contains 200 sentences and almost 550 instances of entities. The three most frequent entity types selected are: (i) disorder, (ii) chemicals and drugs, as well as (iii) procedures. The remaining types are discarded for not having enough instances to train a machine learning model (i.e. less than 100 examples).

Raw Corpus: The Open-Source Parallel Corpus (OPUS) (Tiedemann, 2009) European Medicines Agencies (EMEA) is a set of parallel and unlabeled texts in 22 languages extracted from the EMEA. It consists of documents concerning medicinal products.

The aim of the sampling was to consider the same medicinal products in Spanish and English. The selection of this pair of languages was motivated by their being among the most widely spoken languages (Eberhard, Simons, and Fennig, 2019). A random sampling was performed to choose 16% (i.e. 75 out of 472) medicinal products. All parallel documents available were selected, consisting of at least one hundred thousand sentences for each language — the dataset included $130,782$ sentences for Spanish and $110,403$ sentences for English (see Table 4.1).

We estimated that a sufficiently large number of sentences was $100,000$, and thus, the minimum amount of documents necessary to reach this volume was sampled.

4.2 Results

As a tangible result, the OWLs produced for both languages are available via Zenodo (Moreno and Piad, 2020) and GitHub[1].

The coherence assessment (see Table 3), in both languages, indicates that better results are obtained for the proposed combination of LETO+CARMEN than the original baseline of using only LETO without domain-specific entities. The reviewers assigned a classification of *agree* or *strongly agree* with respect to a random sample of 100 triplets, in 58% of the cases in English, and 62% of the cases in Spanish. This represents an improvement of approximately 21% for the English and 33% for the Spanish language corpora.

In the case of domain growth, Table 4 shows the absolute and relative number of instances detected by the standard LETO sensor (*Actions* and *Standard Entities*) and the ones detected by the CARMEN-based sensor, for both languages. Instances have been divided into *Actions* (which are not treated by CARMEN) and the remaining entities, whereby the ones not automatically tagged by LETO (i.e., not part of the *Standard Entities*) are then passed to CARMEN for fine-grained classification. Both for the English and Spanish corpora, a large fraction of the entities are detected only when CARMEN is used (59.07% and 47.28%, respectively). In the case of actions, a relatively larger percentage is detected for the Spanish language, given that Spanish verbs have a larger variation due to conjugation rules. Applying a standard stemming using NLTK(Bird, Klein, and Loper, 2009) to actions, we find that all 171 English actions have different stems, while in Spanish the 511 actions are grouped into only 277 different stems.

If we represent the stored knowledge as a graph, connecting subjects, verbs and targets, then Figure 3 shows the distribution of degree (number of relations in which each node participates) for each node of graph, i.e., each instance of action, subject or target. The figure shows the result in English and Spanish data, before and after the Knowledge level. In both languages the behavior is very similar. The original graphs (left figures) show a larger number of instances participating in a smaller number of relations, whereas after

[1] https://github.com/knowledge-learning/leto-carmen-ontologies/raw/master/data/results_en.owl

Language	#Sentences	#Words	#Instances
English	200	3,107	539
Spanish	200	3,501	538

Table 1: Number of sentences, words and instances for each parallel language selected from Mantra Gold standard.

	Total		Selected	
Language	#Sentences	#Tokens	#Sentences	#Tokens
English	1.2 million	12.1 million	110,403	1,479,183
Spanish	1.2 million	13.9 million	130,782	2,145,557

Table 2: Number of sentences and words for each parallel language selected from OPUS EMEA corpus, before and after random sampling.

the clustering algorithm is applied, a considerably smaller number of instances participate in more relations (see the difference in the scales of the Y axes). Hence, the clustered graphs are more densely connected.

5 Discussion

Our proposal adapts LETO, a general-purpose knowledge discovery framework, with CARMEN, an entity classification system, to incorporate health-related semantics. The results confirm our objective. Figure 3 shows that clustering connects previously disconnected nodes, forming contextual graphs with over five triplets. This demonstrates the framework's extensibility to specific domains and languages using sensors like CARMEN, which are language- and domain-independent.

Figure 4 displays a subset of the English graph with 50 high-degree nodes, excluding coreferences (e.g., pronouns). Node size reflects connectivity, where larger nodes indicate more relations. The graph highlights key health-related triples such as (patients, have, symptoms), (doctors, treat, patients), and (patient, reduce, symptoms), illustrating the healthcare process.

The analysis of the knowledge discovered ranges from coarse to fine-grained. Actions such as 'include' or 'be' find that symptoms can be: dizziness, rash, nausea, dehydration, somnolence, headache, pain, and fever. This demonstrates that relevant and coarse-grained knowledge is found.

The triplets found fine-grain knowledge, ranging from general concepts, such as disorder, to more specific types such as disease: `(disorders, include, disease)`. An important fact is that true knowledge, specific to the health domain, is available in the final ontology, even if it does not appear explicitly in the evaluation corpus. An example can be the triple (Kinzalkomb, be, combination) that our approach generated. Looking at the summary of the European public assessment report (EPAR) for the medicine Kinzalkomb (available online in [2]), the EPAR says "Kinzalkomb is a medicine that contains two active substances, telmisartan and hydrochlorothiazide". Similarly, one aspect of the knowledge extracted is (fendrix, give, headache). Figure 5 shows three possible sentences that appear in the corpus from which LETO+CARMEN might have extracted this knowledge.

Without extending LETO with CARMEN, true domain-specific knowledge like this would not have been inferred (see domain growth in Table 4). Our approach also enables the conceptualization of semantically related terms, which are simplified or unified by others that appropriately represent them. This is achieved through the clustering process at the Knowledge Level, as shown in Figure 3, which reduces the information volume and summarizes data interpretation.

A key limitation of the entity extraction process is that entities are first detected by LETO and then classified by CARMEN. This restricts the recognition of entities LETO could identify independently, especially for multi-word entities. In this proposal, only multi-word entities classified as PERSON, ORGANIZATION, etc., are recognized. This can be improved by extending CARMEN to handle both entity classification and the initial extraction step, such as by adding a module to detect health-related entity boundaries.

Another limitation arises in the compacting process. Some triplets are semantically incorrect in the health domain because LETO's clustering only considers graph structure, not domain features. Including domain-specific

[2] `https://www.ema.europa.eu/en/medicines/human/EPAR/kinzalkomb`

		Percent		
Language	*Evaluation*	**LETO**	**LETO+CARMEN**	*Improvement*
English	disagree	62,63%	42,00%	- 20,63
	agree	10,10%	11,00%	0,90
	strongly agree	27,27%	**47,00%**	19,73
Spanish	disagree	71,00%	38,00%	- 33,00
	agree	13,00%	21,00%	8,00
	strongly agree	16,00%	**41,00%**	25,00

Table 3: Results in terms of coherence with the original Spanish and English corpora based on 100 triplets.

	English		**Spanish**	
Entity	Instances	% of Total	Instances	% of Total
Actions	171	13.67%	511	29.57%
Standard Entities	341	27.26%	400	23.14%
CARMEN Entities	739	59.07%	817	47.28%
Total	1251		1728	

Table 4: **Domain Growth**. Percentage of entities discovered by the standard LETO sensors and the CARMEN-based sensor for both Spanish and English.

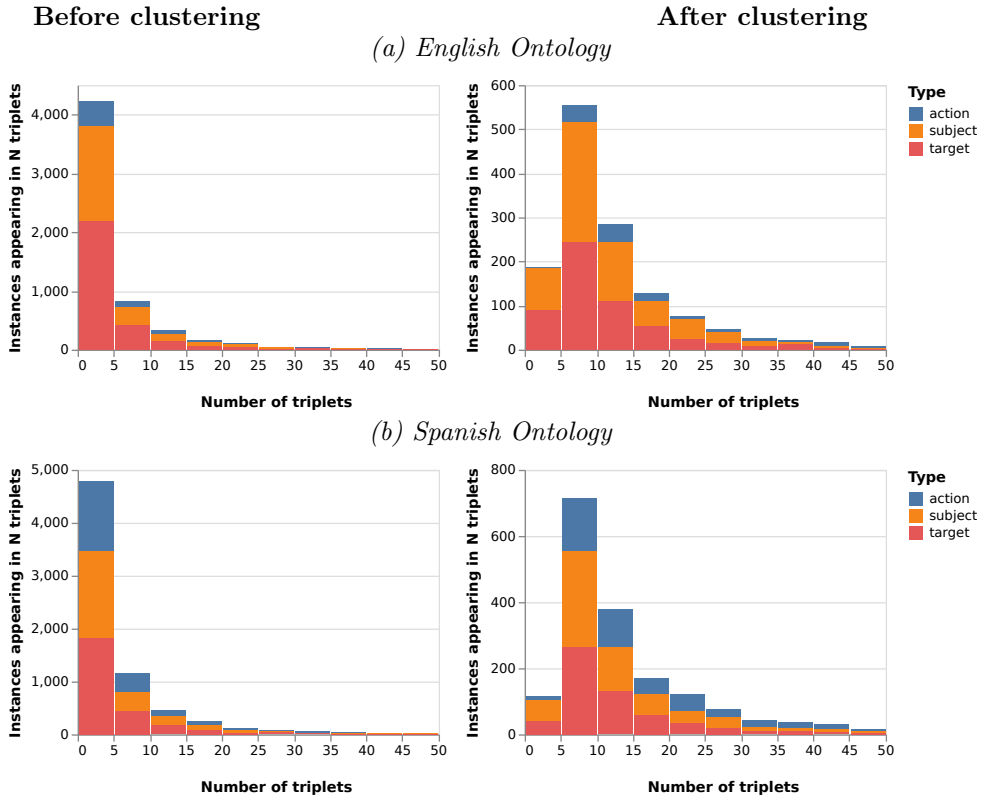

Figure 3: **Reduction Rate.** Histogram of the number of relations per instance in the English (a) and Spanish (b) ontologies, before and after the clustering step (left and right respectively).

features, like fine-grained entity types, could refine clustering, ensuring it groups semantically related entities. Additionally, some nodes with orthographic variations (e.g., patient vs. patients) represent the same concept. This issue arises because no linguistic knowledge is used in compacting. For languages like Spanish, adding stemming or lemmatization in the clustering process would improve the final ontology's connectivity.

Despite these limitations, the achievements are certainly encouraging for the two languages analysed. Besides, we foresee our approach (LETO+CARMEN) to be extremely valuable for languages with scarce resources due to its simplicity and efficacy.

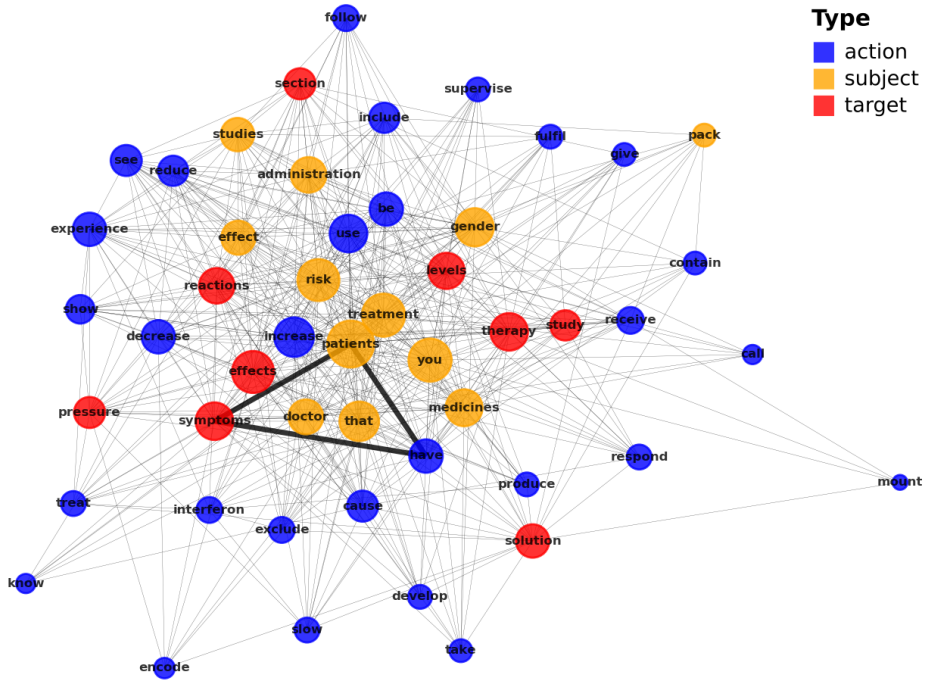

Figure 4: Top 50 instances in the English ontology, node size reflects degree.

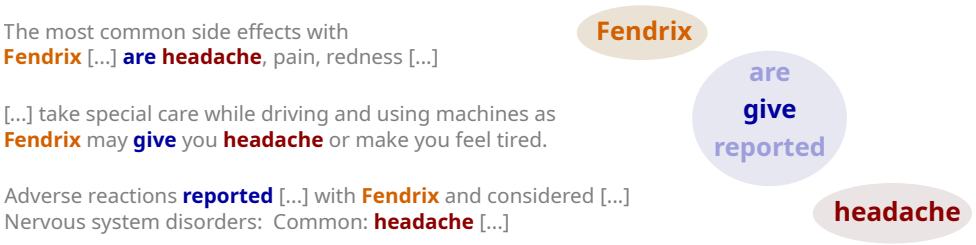

Figure 5: Example phrases and illustrative output by LETO+CARMEN.

6 *Conclusions*

This paper presents a knowledge discovery framework for health technical documents, extending LETO with CARMEN, a multilingual entity classification system, to enhance semantics. Improvements in coherence, domain growth, and reduction are demonstrated for Spanish and English. The approach generates a semantic graph per language, consolidating knowledge and linking insights across documents.

Results show that integrating a domain-specific entity classifier increases both coarse and fine-grained knowledge extraction, enabling the discovery of otherwise undetected health-related relationships. The resulting semantic graph can be applied to large text corpora, supporting decision systems and natural language generation. It aids the medical community in answering complex queries by extracting relevant semantic relations, expanding domain-specific knowledge.

Funding:

This research has been partially supported by the following funding sources: At the national level: COOLANG (PID2021-122263OB-C22); CORTEX (PID2021-123956OB-I00); and CLEART-EXT (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by ERDF A way of making Europe, by the European Union or by the European Union NextGenerationEU/PRTR. At regional level, the Generalitat Valenciana (Conselleria d'Educacio, Investigacio, Cultura i Esport), FEDER and AVI granted funding for NL4DISMIS (CIPROM/2021/21) and EATITALL(INNEST/2023/10).

References

Asim, M. N., M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi. 2018. A survey of ontology learning techniques and applications. *Database*, 2018, 10. bay101.

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

Boag, W., E. Sergeeva, S. Kulshreshtha, P. Szolovits, A. Rumshisky, and T. Naumann. 2018. Cliner 2.0: Accessible and accurate clinical concept extraction. *CoRR*, abs/1803.02245.

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 01.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Brown, E. G., L. Wood, and S. Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117, Feb.

Eberhard, D. M., G. F. Simons, and C. D. Fennig, editors. 2019. *Ethnologue: Languages of the world*. SIL international. Accessed: 2019-09-04.

Estevanell-Valladares, E. L., S. Estevez-Velarde, A. Piad-Morffis, Y. Gutierrez, A. Montoyo, R. Muñoz, and Y. Almeida Cruz. 2021. Knowledge discovery in COVID-19 research literature. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 402–410, Held Online, September. INCOMA Ltd.

Estevez-Velarde, S., Y. Gutierrez, A. Montoyo, A. Piad-Morffis, R. Munoz, and Y. Almeida-Cruz. 2018. Gathering object interactions as semantic knowledge. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 363–369. The Steering Committee of The World Congress in Computer Science.

Estevez-Velarde, S., A. Montoyo, Y. Almeida-Cruz, Y. Gutiérrez, A. Piad-Morffis, and R. Muñoz. 2019. Demo application for LETO: Learning engine through ontologies. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 276–284, Varna, Bulgaria, September. INCOMA Ltd.

Explosion. 2019. Spacy [online]. Accessed: 2019-05-23.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, et al. 1996. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88.

Feldman, R. and J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, New York.

Frey, B. J. and D. Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Friedman, C., L. Shagina, Y. Lussier, and G. Hripcsak. 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 09.

Galiano, S., R. Muñoz, Y. Gutiérrez, A. Montoyo, J. I. Abreu, and L. A. Ureña. 2023. T2kg: Transforming multimodal document to knowledge graph. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 385–391.

IHTSDO. 2017. *SNOMED Clinical Terms Starter Guide*. 2017 edition. (Last accessed: 31/05/2019).

Konys, A. 2018. Towards knowledge handling in ontology-based information extraction systems. *Procedia Computer Science*, 126:2208 – 2218. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

Kors, J. A., S. Clematide, S. A. Akhondi, E. M. Van Mulligen, and D. Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: The Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Lopes, L. and R. Vieira. 2015. Building and Applying Profiles Through Term Extraction. In *X Brazilian Symposium in Information and Human Language Technology*, pages 91–100, Natal, Brazil.

McCray, A. T., A. Burgun, and O. Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.

Moreno, I., E. Boldrini, P. Moreda, and M. Romá-Ferri. 2017. Drugsemantics: A corpus for named entity recognition in spanish summaries of product characteristics. *Journal of Biomedical Informatics*, 72:8–22.

Moreno, I. and A. Piad. 2020. knowledge-learning/leto-carmen-ontologies: English and Spanish OWL files, March.

Moreno, I., M. Romá-Ferri, and P. Moreda. 2019. Carmen: An entity typing system applied to medical corpora that is language-independent and entity type set independent. *Artificial Intelligence in Medicine*. (submitted).

Patrick, J. D., D. H. M. Nguyen, Y. Wang, and M. Li. 2011. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association*, 18(5):574–579, 07.

Roberts, A. 2017. Language, structure, and reuse in the electronic health record. *AMA journal of ethics*, 19(3):281–288.

Rogers, F. B. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51(1):114–116.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 09.

Tiedemann, J. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, pages 237–248.

Viani, N., C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, and L. Sacchi. 2018. Information extraction from italian medical reports: An ontology-driven approach. *International Journal of Medical Informatics*, 111:140 – 148.

Vítores, D. F. 2017. El español: Una lengua viva. informe 2017 [spanish: a living language. report 2017]. Technical report, Instituto Cervantes.

Wong, W., W. Liu, and M. Bennamoun. 2012. Ontology learning from text: A look back and into the future. *ACM Comput. Surv.*, 44(4), September.