A corpus to support eHealth Knowledge Discovery technologies

Alejandro Piad-Morffis[a,*], Yoan Gutiérrez[b], Rafael Muñoz[b]

[a]*Department of Artificial Intelligence and Computing Systems, University of Havana*
San Lázaro y L. Edificio Felipe Poey. Plaza de la Revolución, Havana (Cuba)
[b]*Department of Software and Computing Systems, University of Alicante*
Carretera San Vicente del Raspeig s/n, 03690, Alicante (Spain)

Abstract

This paper presents and describes eHealth-KD corpus. The corpus is a collection of 1173 Spanish health-related sentences manually annotated with a general semantic structure that captures most of the content, without resorting to domain-specific labels. The semantic representation is first defined and illustrated with example sentences from the corpus. Next, the paper summarizes the process of annotation and provides key metrics of the corpus. Finally, three baseline implementations, which are supported by machine learning models, were designed to consider the complexity of learning the corpus semantics. The resulting corpus was used as an evaluation scenario in TASS 2018 [1] and the findings obtained by participants are discussed. The eHealth-KD corpus provides the first step in the design of a general-purpose semantic framework that can be used to extract knowledge from a variety of domains.

Keywords: Corpus, Subject-Verb-Object, Knowledge Discovery, Spanish, eHealth
2010 MSC: 68T35, 68T50

1. Introduction

The accelerated growth of the Internet has resulted in a massive collection of scientific texts that are available online. Several bibliographical databases exist, grouping academic texts from different domains, such `Arxiv.org`[1] and `Medline`[2], which are two of the largest repositories, containing a vast amount of information that can be used by the scientific community. However, its large size makes it impossible for human researchers to efficiently find useful results, definitions, or facts. Even with the use of specialized search engines (such as Google Scholar), it is complicated to find relevant information in domain-specific documents. This is due in part to the lack of a unified semantic structure in these documents, which are written in natural language.

To provide more fine-grained search results, documents can be processed to extract the relevant semantic entities and facts mentioned. The task of automatically discovering semantic knowledge from text is covered by research areas such as ontology learning [2] and learning by

*Corresponding author: `apiad@matcom.uh.cu` (+34) 965903400 ext. 2961.
Email addresses: `apiad@matcom.uh.cu` (Alejandro Piad-Morffis), `ygutierrez@dlsi.ua.es` (Yoan Gutiérrez), `rafael@dlsi.ua.es` (Rafael Muñoz)

[1] https://arxiv.org
[2] https://medline.gov

reading [3], whose purpose is to build semantic networks that capture the knowledge present in large collections of text. These semantic networks enable the use of search engines that provide an analysis beyond the textual content's relevance, by exploiting the semantic structure of the network. In this context, processing health textual contents has attracted great interest [4], motivated by the large number of medical documents published yearly.

Several approaches exist for building semantic representations of knowledge. In many cases, these representations use a domain-specific conceptualization. Although this provides a more specialized representation, it makes these approaches harder to apply to a broad range of domains. Alternatively, a general purpose conceptualization could be used, which is able to represent entities and facts from multiple knowledge domains. Such conceptualization should be general enough so as to accommodate many different domains, but still to provide a degree of expressiveness necessary for knowledge mining tasks. One possible conceptualization is using Subject-Action-Target triplets [5]. This structure has proven to be useful for representing knowledge in both specific domains such as movie reviews [5] or sentiment mining [6] and in general domain ontology learning [7]. Furthermore, Subject-Action-Target triplets automatically extracted from text can be later linked to domain-specific relations through the use of semantic networks. As an example, the *SemRep* system [8] extracts Subject-Predicate-Object triplets from natural eHealth texts. The predicates are linked to specific relations in the UMLS [9] semantic network.

Recent work in the development of Teleologies [10] suggests that Action-Subject-Target triplets can be the base for general purpose conceptualizations across many different domains, since this triplet allows the capture of interactions between objects through the actions they perform on each other. A small set of semantic relations, such as *hyponomy* and *holonomy* can provide additional semantic structure to theAMR representation. These "general" relations are common in most knowledge bases, regardless of domain, such as WordNet [11], DBPedia [12], and ConceptNet [13]. Other possible conceptualizations allow the capture of semantics of natural language, such as Abstract Meaning Representation (AMR) [14]. Despite the superior representational power of AMR over simple structures such as Action-Subject-Target triplets and basic semantic relations, the annotation process for AMR is considerably more complex both for humans and automated techniques.

Building corpora annotated with the Action-Subject-Target structure is the first step towards the design of systems that can automatically extract these annotations. Several corpora exist in the literature, annotated with a variety of different schemes, such as CLEF [15], Yago [16] and Emotinet [6]. However, most of these resources are annotated with domain-specific conceptualizations that are difficult to extend to different knowledge domains. This paper presents a general purpose conceptualization and an example corpus[3] annotated with such conceptualization, which demonstrates its ability to represent a wide variety of topics in a semantically rich structure. Furthermore, a set of baseline implementations of machine learning techniques for automatically annotating similar sentences are presented[4]. Based on these resources, an ongoing online evaluation is available for researchers[5].

The paper is organized as follows. Section 2 presents a set of relevant corpora that share familiar characteristics with the proposed eHealth-KD corpus. In section 3 the semantic structure of the corpus is defined and justified. Subsequently, section 4 presents the statistics of the corpus,

[3]https://github.com/knowledge-learning/ehealth-kd

[4]https://github.com/knowledge-learning/ehealth-kd/tree/master/baseline

[5]https://competitions.codalab.org/competitions/18188

describes the annotation process (section 4.1) and presents the main evaluation metrics regarding the corpus quality (section 4.2). In section 5 we present a methodology for evaluating automated annotation systems trained on the eHealth-KD corpus and a discussion of current approaches. Finally, section 6 presents the main discussions and insights from this research, and section 7 the final considerations.

2. Related Corpora

This section analyses different corpora relevant to the domain of knowledge extraction in medical texts, as well as other general corpora with a semantic annotation similar to the one presented in this paper. The literature exhibits a large corpus in the medical domain, though few corpora exist that present a general semantic structure suitable for multiple domains. However, this section focuses on the subset that is the most similar with the eHealth-KD corpus, either in terms of content or semantic structure. Hence, even though the largest corpora available in both domains are written in English, we over-sample Spanish corpora in this comparison, since the eHealth-KD corpus contains Spanish documents. Tables 1 and 2 present a summary of the characteristics of the studied corpora.

Corpus	**Drug Semantic**	**Ixa MedGS**	**CLEF**	**DDI**	**BARR2**
Doc. Type	Product summaries	Discharge summaries	Clinical documents	Abstracts	Clinical case studies
Annotation Type	Manual	Auto/Manual check	Manual	Auto/Manual check	Manual
Annotators	Experts	Experts	Experts & Non-experts	Expert	Experts
Schema	Medical entities	Medical entities	Medical entities	Medical entities	Medical abbreviations
Language	Spanish	Spanish	English	English	Spanish
Documents	5 (16%)	75 (0.01%)	150 (0.27%)	1025 (100%)	648(20%)
Origin	AEMPS	Galdacao-Usansolo Hospital	Royal Madersen Hospital	Medline, Drug Bank	PubMed, IBCECS & SciELO

Table 1: Summary of related corpora annotated with domain-specific entities for the health domain. Percentage values for **Documents** indicate how many of the original documents were actually annotated, as reported by the original authors.

Most health-related corpora are annotated using self-defined health related entities relevant to the task at hand. Of these, arguably one of the most used is the CLEF corpus [15]. This corpus contains 150 English clinical documents, manually annotated by a team of experts (clinical and biologists) and non-experts. In contrast, the DDI corpus [17], which contains 1025 English documents from Medline was pre-annotated automatically, and then manually checked by domain experts (Pharmacists). Similar corpora in Spanish language exist. The Drug Semantic corpus [18] is an example, where domain experts (Registered Nurses and students) manually annotated Spanish summaries of product characteristics. Likewise, the BARR2 [19] corpus contains manually annotated abbreviation-definition pairs in Spanish clinical papers extracted from bibliographic databases. On the other hand, the Ixa MedGS corpus [20] was pre-annotated automatically and then manually checked by domain experts in Pharmacology. An interesting alternative is the

Corpus	Bio AMR	Yago	Emotinet	eHealth-KD
Doc. Type	Sentences	Sentences	Posts	Sentences
Annotation Type	Manual	Automatic	Manual	Manual
Annotators	Non-experts	Non-experts	Non-experts	Non-experts
Schema	AMR	SPO	SAOE	SAT+R
Language	English	English	Spanish & English & Italian	Spanish
Documents	6542	—		1173(11.8%)
Origin	PubMed	Wikipedia, WordNet	Blog	Medline Spanish XML

Table 2: Summary of related corpora annotated with a general-purpose schema, or not specific to the health domain. Percentage values for **Documents** indicate how many of the original documents were actually annotated, as reported by the original authors. SPO: Subject, Predicate, Object triplets; SAT+R: Subject, Action, Target triplets and additional Relations (see section 3); SAOE: Subject, Action, Object, Emotion tuples.

Bio AMR corpus [21], which contains AMR annotations of several medical documents, hence combining a general purpose annotation schema in a specific domain.

In the context of general domain knowledge, one of the most relevant resources for our research is YAGO [16]. It consists of a large knowledge base automatically extracted from Wikipedia, WordNet, and other sources. Since YAGO is intended to represent general domain knowledge, its semantic structure is defined in terms of fact triples, in the spirit of RDF and other ontological representations. In contrast, the Emotinet knowledge base [6] is oriented towards a specific domain (emotions), and is built from the manual annotation of blog entries, using a general semantic structure that links entities, actions, and emotions. Although Emotinet is designed for a particular domain, its structure is rather general, in the sense that it can readily represent any type of event or action performed by entities.

As Table 1 shows, the type of documents used is highly variable, which provokes large differences in terms of the length of documents, structure of discourse and vocabulary. An interesting characteristic is the type of annotation, either manual, pre-automated with expert review, or fully automated. Although recent research shows an increasing tendency towards pre-automated or fully automated annotation, manual annotation is still regarded as more reliable.

Health related corpora are usually annotated by experts with a domain-specific semantic structure, such as entities related to diseases, drugs, genes, or treatments. Given the complexity of the concepts in the medical domain, annotators usually include medical doctors or other specialists of the medical domain. In these resources, very few general-purpose natural language features are used. This provides a greater detail of semantic information, since the entities and relations are relevant for the domain at hand. However, in the same sense, it might discard important information in the text which cannot be represented with the structure defined. This may or may not be an issue for a specific line of research. In our case, we consider it important to extract as much knowledge as possible from each source. In contrast, general purpose corpora or knowledge bases are usually annotated by non-experts with a semantic structure designed to represent as much knowledge as possible. This strategy tends to increase recall (a larger amount of facts is extracted) but it might extract irrelevant or incorrect facts. In these cases, the annotation schema relies largely on natural language semantics, such as Subject-Predicate-Object triples.

The trend of representing knowledge with a general structure has been aided by recent advances in Teleologies [10] that provide a theoretical framework for representing general purpose

facts using a small set of concepts (objects, actions and functions). In contrast with Abstract Meaning Representation (AMR), the Teleologies framework is not specifically aimed at natural language understanding, but at representing the semantics of a general knowledge domain. This type of framework is less dependent on the linguistic characteristics of a specific language. The Subject-Action-Target structure defined in this paper is based on a simplification of the Teleologies conceptualization, applied to the domain of medical texts. However, inspired by general purpose knowledge bases, we also include a few specific semantic relations that are broadly used in general purpose ontologies and semantic networks. This combination (i.e. SAT+R, see Section 3) makes the annotation schema used in eHealth-KD novel.

3. Semantic structure of the eHealth-KD corpus

In designing the semantic structure of the presented corpus, two general purpose conceptualizations are analyzed, the Abstract Meaning Representation (AMR) [14] and the Teleologies framework [10]. In our proposal, named SAT+R (Subject, Action, Target and Relations), to capture the fundamental semantics of a broad range of text, we propose a simplified version of the Teleologies conceptualization. This consists of the identification of two key elements in the text: **Concepts** and **Actions**, that roughly maps to the layers of *Objects* and *Actions* in Teleologies. However, the *Functions* layer of Teleologies has not been considered at this stage.

Furthermore, since the purpose of our conceptualization is to support knowledge discovery technologies, we also draw inspiration from general-purpose knowledge bases and ontologies such as DBPedia and ConceptNet. Based on the structure of these knowledge bases and in line with semantic annotations (i.e. `HYPONYM-OF`, `SYNONYM-OF`) promoted in shared campaigns like SemEval 2017 Task 10 [22], we define four general purpose semantic relations: **is-a**, **same-as**, **part-of** and **property-of**. These relations allow to directly represent important general-purpose semantics that can appear in many different textual forms, resulting in a more compact and normalized annotation. Our annotation schema is thus an hybrid schema that attempts to capture as much information as possible through the use of Subject-Action-Target triplets, while also specifically recognizing important semantic relations that appear in most general purpose knowledge bases.

Intuitively Concepts represent actors or entities which are relevant in a domain, while Actions are a particular type of Concept which represent how other Concepts interact with each other. Actions and Concepts can be linked by two types of relations: **Subject** and **Target**, which describe the main roles that a Concept can perform. Figure 1 shows an example annotation of a small set of sentences with the corresponding labels and relations.

The Concepts are those key phrases that are able to represent objects and other entities presumed to be of interest for some particular purpose. It is possible to represent simple and complex Concepts. Simple Concepts just represent singular entries like "*asma*" (*asthma*) or multi-words like "*vías respiratorias*" (*respiratory tract*), etc. Complex Concepts are explained below in this section. An Action is a type of Concept which provokes a modification of another Concept, commonly represented by verbs or phrases that include verbs, but in some cases, it can be represented by a non-verb.

As explained, Concepts can play two different roles in an Action:

Subject : A role identifying the actor that performs the indicated action. For example, in "*el asma afecta las vías respiratorias*" (*asthma affects the respiratory tract*), the Con-

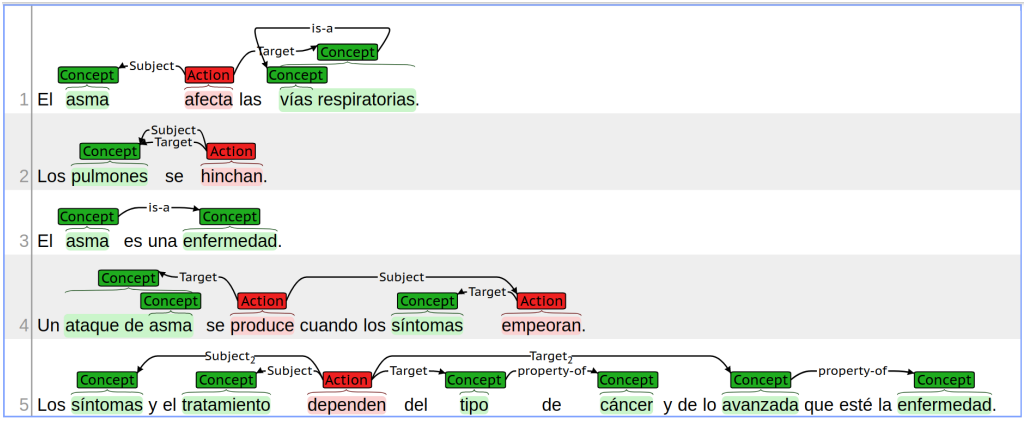

Figure 1: Example annotation of a small set of documents.

cept "*asma*" is what performs the action "*afecta*". It can be said that a subject plays the producer/actor role in this relationship.

Target : A role identifying the actor that receives the effect of the indicated action. For example, in "*el asma afecta las vías respiratorias*", the Concept "*vías respiratorias*" receives the effect of the action "*afecta*". It can be said that a target plays the recipient role in this relationship.

Although Actions can have both a Subject and a Target, sometimes the subject of an action is hidden, or non-existent. Such is the case of actions represented in Spanish by infinitive verbs, for example in the sentence "*Diagnosticar el cáncer es difícil*" (*diagnosing cancer is difficult*). In this example the Action "*Diagnosticar*" only has a target, the Concept "*cáncer*", since who performs the action is not specified in the sentence.

In other examples, the target can either be missing, or be the same as the subject. Such is the case of actions represented by reflexive verbs, for example in the phrase "*...los pulmones se hinchan*" (*the lungs swell*). In this example the Action "*hinchan*" (*swell*) has the same subject and target, the Concept "*pulmones*" (*lungs*). This means that the subject and the target refer both to same concept. As observed in the previous example, an Action can have more than one Subject and/or Target, if the same fragment of text is used to denote multiple occurrences of said action.

Besides these roles, we define 4 additional semantic relations among Concepts, summarized in the following list. These relations are preferred over generic Subject-Action-Target when possible, since they have more a specific semantic meaning that need not be inferred from the surface text of an Action annotation.

is-a: indicating that the first Concept is a sub-type, or more concrete expression of the second Concept. For example, "*asma*" is-a "*enfermedad*" (*disease*); or "*vías respiratorias*" is-a "*vías*".

part-of: indicating that the first Concept is a constituent part or component of the second Concept, such as in "*pulmones*" part-of "*cuerpo humano*" (*human body*).

property-of: indicating that the first Concept defines any property or variable characteristic of the second Concept, such as in "*avanzada*" (*advanced*) property-of "*enfermedad*".

same-as: for indicating a concept is unambiguously the same as another concept. For example, in "*... el Síndrome de Inmunodeficiencia Adquirida (SIDA)...*", the Concept "*Síndrome de Inmunodeficiencia Adquirida*" (*Adquired Inmunodeficiency Syndrome*) is the same-as "*SIDA*" (*AIDS*).

Even though Actions are conceptually a special type of Concept, we do not currently consider the previous 4 semantic relations between Actions. This issue will be dealt with in future versions of the corpus.

Complex concepts are represented by tuples, e.g <Subject,Action,Target> (or any variant where target or subject can be missing), in which the Action constitutes its core. Therefore, sometimes the subject or target can be another type of Action, which represents a complex concept. For example, in the sentence "*Un ataque de asma se produce cuando los síntomas empeoran*" (*An asthma attack occurs when symptoms get worse*), a complex concept is "*síntomas empeoran*" where "*síntomas*" (*symptoms*) is a Concept and "*empeoran*" (*get worse*) is an Action; "*síntomas empeoran*" is the act of symptoms getting worse. Therefore, "*síntomas empeoran*" can be linked to the Action "*produce*" (*occurs*). The Action "*produce*" in this sentence is performed by this complex concept, i.e., it is not the symptoms that cause the asthma attack, but rather it is the act of the symptoms getting worse that causes the asthma attack.

4. Corpus description

The corpus was built with an XML file taken from Medline on January 9th, 2018[6]. The exact file processed is an XML dump from January 9th, 2018. The original file is not available online at the moment of writing, but we provide a copy for reproducibility purposes[7]. This file contains 2026 entries in the Spanish language from several health-related topics. A selection of these files was annotated according to the semantic interpretation of each sentence, based on the Action-Subject-Target structure previously defined. Each entry was parsed, split by sentences, and then some additional cleanup was performed, such as removing copyright or authorship notes, removing sentences ending in "*?*" or "*!*" or sentences with less than 5 words. Additionally, HTML-specific markup, such as lists or anchors, was also removed. Finally, there are 9956 sentences, split across 41 files, grouped by topic. All relevant software and data used in to build the corpus is available online[8].

Using this pool of sentences, an annotation workflow was implemented to manually tag the relevant entities and relations described in Section 3. This annotation process is described in detail in Section 4.1. After this annotation process a total of 1173 sentences were obtained and split across 4 collections, described below. Table 3 summarizes the statistics of the final corpus.

The trial collection contains 29 sentences. This collection was created before starting the annotation process, with the purpose of reaching a common consensus among the annotators. The trial collection is a summary of all the possible annotation patterns that appear in the text. From this collection, an annotation guide is created to aid the annotators.

[6]https://medlineplus.gov/xml.html

[7]https://github.com/knowledge-learning/ehealthkd-2018-dev/blob/master/scripts/data/mplus_topics_2018-01-09.xml?raw=true

[8]https://github.com/knowledge-learning/ehealthkd-2018-dev

Metric	Overall	Trial	Training	Develop	Test
Files	11	1	6	1	3
Sentences	1173	29	559	285	300
Annotations	13113	254	5976	3573	3310
Entities	7188	145	3280	1958	1805
- Concepts	5366	106	2431	1524	1305
- Actions	1822	39	849	434	500
Roles	3586	71	1684	843	988
- subject	1466	33	693	339	401
- target	2120	38	991	504	587
Relations	2339	38	1012	772	517
- is-a	1057	18	434	370	235
- part-of	393	3	149	145	96
- property-of	836	15	399	244	178
- same-as	53	2	30	13	8

Table 3: Statistics of the eHealth-KD v1.0 corpus.

The training collection contains 559 sentences split across 6 files. Each file contains sentences related to one topic. Therefore, within one file there can be repetitions of common concepts and actions (i.e., diseases are mentioned more than once), but between 2 different files there is significantly lower degree of intersection. This collection is the main resource for training or fitting a learning system.

The development collection contains 285 sentences in a single file. These sentences are intermixed from different topics and shuffled. This makes this collection suitable for hyperparameter tuning, model selection and model validation, and to help reduce over-fitting.

The test collection contains 300 sentences split across 3 files of 100 sentences each. These sentences come from different topics. Furthermore, they were carefully selected so that approximately 50% of the annotations (entities and relation pairs) are syntactically identical to annotations in the training collection, although no sentence is exactly similar to any sentence in the training collection. The other 50% of the annotations are syntactically different to any annotation in the training collection. Hence, systems which are only dependent on textual features (i.e., lexemes) are expected to perform below a 50% accuracy. In order to achieve better performance, additional semantic features must be considered.

4.1. The annotation process

To ensure a consistent annotation across all the corpus, the annotation process was split in four stages. All annotations were performed using the Brat annotation tool [23]. Among the annotators, there are two groups: **expert** annotators (3) and **non-expert** annotators (12). The expert annotators are researchers specialized in semantic analysis of natural text (PhD and PhD students) and the non-expert annotators are computer science students and post-grad students and professors, all native Spanish speakers. The expert annotators created the trial collection and the annotation guide (stage 1) and performed the final normalization (stage 4). The non-expert annotators were involved in the annotation process (stages 2 and 3). Figure 2 shows a schematic representation of the whole process.

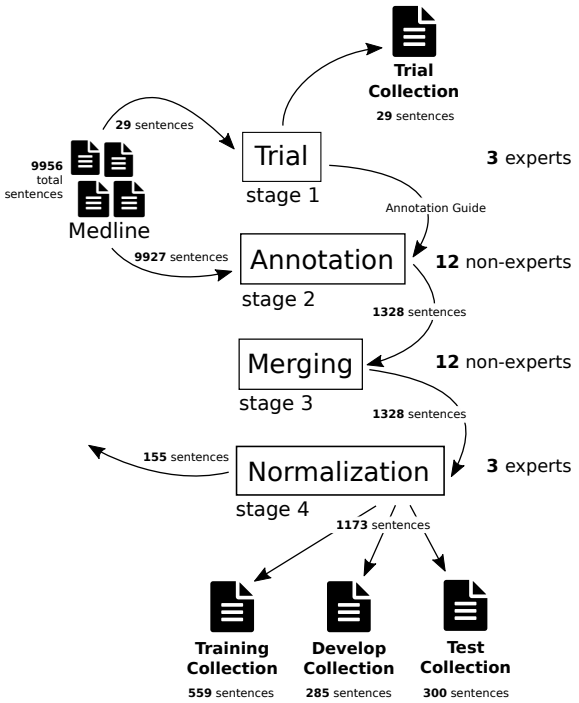

Figure 2: Schematic representation of the annotation process.

Stage 1. In the first stage, 29 random sentences were independently annotated by 3 expert annotators. Afterwards, all three versions of the annotations were compared and normalized, to achieve consensus. The result of this process is the trial collection. From this trial collection an annotation guide was created. The annotation guide consists of a subset of these sentences, manually selected, plus other artificial sentences, specifically designed to highlight key aspects of the annotation. Specific attention was given to the disambiguation of possible contradictions, such as when to use Actions versus semantic relations (**is-a**, etc.) and how to differentiate **property-of** and **part-of**. The annotation guide and the trial sentences were distributed to the rest of the annotators as reference.

Stage 2. In the second stage, all the files were split and assigned to 6 teams of 2 non-expert annotators. In each team, both annotators independently annotated the assigned files. In this stage, a total of 1328 different sentences were annotated by two different annotators. During the annotation phase the exact number of sentences annotated by 6 persons in a continuous period of 30 minutes was measured. This resulted in an average annotation time of 70 seconds per sentence.

Stage 3. In the third stage, an automatic merging process was then performed, which identified for each sentence the matching and conflicting annotations. The result of this process was a collection of "*merge*" files with the union of their respective annotations, clearly identifying which annotations were conflicting and which were exactly matched between both versions. At this point, 47% of the annotations were identified as conflicting. However, a manual review demonstrated that most of the conflicts were simple syntactical differences that did not indicate a semantic disagreement, such as phrase boundary errors (e.g., including the final stop or a comma in a key phrase).

Each "*merge*" file was then manually inspected by a different non-expert annotator (someone not involved in the second stage with that specific file). When the differences were deemed as clearly fixable, the annotator would manually change them. When the differences indicated conflicting semantic meanings, the annotator had to decide which of the conflicting annotations was correct, or give a third and decisive variant. The result of this stage was a collection of "*normalized*" files.

Stage 4. Finally, in the fourth and final stage, 3 expert annotators (not involved in the previous stages) analyzed each of the "*normalized*" files independently. If any of the annotators did not agree with a sentence, then this sentence would be publicly discussed until consensus was reached or the sentence would be discarded if no consensus could be achieved. The result of this process was 1144 sentences that were split into the training, development and test collections. Hence, each of the final sentences was reviewed independently by three annotators reaching mutual consensus.

These sentences were grouped into 12 different files according to topics. Of these, 6 files were selected for the training collection. The remaining 6 files were concatenated, and 300 sentences were randomly selected for the testing collection, with a careful sampling procedure to ensure that roughly 50% of the annotations coincided with annotations in the training context, and the remaining 50% did not belong to the same context. The remaining sentences were selected for the development collection.

4.2. Corpus quality metrics

Evaluating the eHealth-KD corpus quality is a complex task, given the manual nature of the annotation, and the characteristics of the domain. Hence, we propose a set of different evaluation metrics that focus on key aspects of the corpus. The first aspect to consider is the quality of the annotation. In this respect, inter-annotator agreement is an important metric to consider. Since the eHealth-KD corpus has been annotated in several stages, and at each stage different annotators were involved, it is possible to measure how the different versions of each annotation evolved during the process, and degree of agreement reached. Cohen's Kappa [24] is a common choice when evaluating inter-annotators agreement. However, this metric applies a binary decision to the final inclusion of each annotation, whereas the eHealth-KD corpus allows for the annotation of text spans and partial matches. Moreover, when large segments of text are not annotated –e.g., all the stopwords, determinants, connectors, and similar lexical elements which are not part of a Concept or Action–, the degree of agreement between annotation versions may be overestimated by Kappa. To account for these factors, we propose to use a metric that considers partial matches and doesn't include the non-annotated portions of text. This metric was inspired by the evaluation criteria designed for the Drug Semantics corpus [18] but has been modified to the specifics of the eHealth-KD corpus.

First, we define a single $G_{class} \in [0, 1]$ metric between overlapping annotations A and B (eq. 1) *of the same class* from different annotators, where *class* can be either *action* or *concept*. Hence, annotations of different classes are counted separately. This metric reaches its maximum value $G_{class}(a_i, a_j; b_i, b_j) = 1$ if the text spans of annotation (a_i, a_j) and (b_i, b_j) overlap exactly (i.e. $a_i = b_i$ and $a_j = b_j$), and a correspondingly smaller value for partial overlap. In the case that either annotation A or B do not have a corresponding overlap annotation, the G value is defined as 0, which is equivalent to defining $b_i = b_j$ (or alternatively $a_i = a_j$).

$$G_{class}\left(a_i, a_j; b_i, b_j\right) = \frac{min(a_j, b_j) - max(b_i, a_i)}{max(a_j, b_j) - min(b_i, a_i)} \tag{1}$$

When a multi-word Concept includes a single word concept such as in "*ataque de asma*" and "*asma*", there are several possible combinations. In this case, both sets of annotations are sorted by size, and then matched accordingly (i.e., the largest of A with the largest of B, and so on). This simple heuristic can underestimate agreement if, for example, annotator B doesn't select the multi-word concept but only the single word, but nevertheless is a sensible solution to avoid computing all possible permutations and selecting the one with largest agreement. The set of all G_{class} values for the whole dataset (of the same class) are micro averaged, to provide a single μG_{class} for each type of annotation (*concept* or *action*) (eq 2).

$$\mu G_{class} = \frac{1}{n} \sum_{k=1}^{n} G_{class}\left(a_i^k, a_j^k; b_i^k, b_j^k\right) \tag{2}$$

Finally, a macro average of G_{action} and $G_{subject}$ provides a single quality value for the task of annotating Concepts and Actions (eq. 3).

$$\mu G = \frac{\mu G_{action} + \mu G_{concept}}{2} \tag{3}$$

Besides the annotation of Concepts and Actions, the other relevant element in the corpus are the 4 semantic relations and the 2 semantic roles defined in Section 3. Each of these 6 relations involves two Concepts and/or Actions of the same sentence. To account for possible

Metric	Stg 2	Stg 3 (A)	Stg 3 (B)	Stg 3	Stg 4
Dismissed (%)	0 (0)	150 (0.11)			5 (0)
Annotations $\mu G_{concept}$	0.50	0.70	0.75	0.73	0.85
Annotations μG_{action}	0.44	0.66	0.70	0.68	0.87
Annotations μG	0.47	0.68	0.73	0.71	0.86
Relations μH	0.24	0.47	0.56	0.52	0.73
Overall F_1	0.32	0.56	0.63	0.60	0.79

Table 4: Summary of the evaluation metrics in each stage.

disagreement in two annotations, we define a metric $H_{rel} \in \mathbb{N}$ equal to the number of coincident relations agreed by both annotators. Likewise, we define R_{rel} as the number of instances of relation pairs *rel* in the union of A and B. Hence, we can define the agreement of annotators A and B for one particular relation as the quotient of these two values. However, since the number of instances for each relation is very different, we propose to micro-average across all relation types, instead of computing a separated agreement on each. Hence, Equation 4 defines an averaged μH metric which considers all types of relations at once.

$$\mu H = \frac{H_{subject} + H_{target} + H_{is-a} + H_{same-as} + H_{part-of} + H_{property-of}}{R_{subject} + R_{target} + R_{is-a} + R_{same-as} + R_{part-of} + R_{property-of}} \tag{4}$$

Finally, a harmonic mean F_1 between μG and μH is computed to provide a single value for corpus quality (eq 5).

$$F_1 = \frac{2 \cdot \mu G \cdot \mu H}{\mu G + \mu H} \tag{5}$$

The F_1 metric is computed for each relevant stage of the annotation in the following manner:

(a) In **Stage 2** between each pair of annotators that tagged the same set of sentences.
(b) In **Stage 3** between each version from the previous stage and the curated result produced by the human annotator.
(c) In **Stage 4** between the result of the previous stage and the result of the normalization by the three human experts.

Table 4 shows the result of the evaluation process of all stages, including the computed F_1 for each stage, the component μG and μH, and other basic metrics of interest, such as number of exact matches and number of missing annotations. Additionally, in each stage we report the number of sentences that were dismissed either because the annotators didn't produce any tag or because the human experts (reviewers) decided the sentence was ambiguous.

As expected, the overall agreement increases with each stage. The final version has an aggregated $F_1 = 0.79$, which is considered adequate according to the Drug Semantics [18] evaluation methodology from which this evaluation was derived. In general, the agreement regarding the annotation of elements (Concepts and Actions) is higher than the annotation of relations. This is consistent with the perceived complexity of annotating both kinds of elements. Intuitively, we expect most annotators to agree on whether a key phrase is an object or an action. There seems to be seldom disagreement on this problem. As for determining the key phrases, there is a greater degree of disagreement, because some annotators fail to detect multi-word phrases.

The largest disagreement occurs in the annotation of relations, particularly in differentiating **is-a** patterns with **property-of**. For example, in the phrase "*...un profesional con licencia...*" one annotator selects "*profesional*" and "*licencia*" as Concepts, related by a **property-of**, while another annotator considers "*profesional*" as an **is-a** of "*profesional con licencia*". Since both make sense semantically, the correct option is subject to the interpretation of the annotator. In these cases, we prefer to annotate the **property-of** variant, unless there exists a relation *with* the larger concept. In the previous example, if there is an action to relate with the concept "*profesional con licencia*", then, and only then, we prefer this annotation.

5. Results of the TASS 2018 eHealth-KD challenge

The corpus presented in this paper was considered as the evaluation scenario for the shared Task 3: "eHealth Knowledge Discovery" in the TASS 2018 Workshop [1]. Participants were given access to the training and development collection gold files, but only the input files for the test collection, and were asked to submit the corresponding outputs.

In order to evaluate the semantic extraction performance of participant systems we proposed to use a standard F_1 metric. However, since there are several different annotations with varying degrees of complexity, we subdivided the overall task into smaller subtasks, that can be evaluated both independently and jointly. This provides a more fine-grained evaluation. Details about the evaluation metrics are provided in the Appendix section.

These subtasks follow a workflow for tackling these problem that is based on our own experience with similar problems and corpora. The tasks are defined as follows:

Subtask A - Identification of key phrases : In this task the only concern is to identify which word n-grams are potential key phrases, either concepts or actions.

Subtask B - Classification of key phrases : In this task, each of the previously identified key phrases is assigned a label, either **Concept** or **Action**.

Subtask C - Discovery of relations : In this final task each pair of entities classified in task B is assigned one or more of the *semantic relations* defined in section 3.

The competition was organized as a set of 3 different evaluation scenarios, each using a different subset (100 sentences) of the test collection. In the first scenario, only input files were given, and participants would provide output for tasks A, B and C. In scenario 2, output files for task A were also provided, and in scenario 3 the outputs for both tasks A and B were given. This setup was designed to evaluate each task both jointly and independently. The results of the competition are discussed in greater detail in the TASS 2018 Overview [1]. For each of these subtasks, the corpus provides gold output files formatted accordingly.

Three baseline techniques are provided for comparative purposes. The first technique is based on simple textual matching, and the two remaining techniques are based on logistic regression and decision trees with simple syntactic and semantic features (e.g., POS-tags and Word2Vec representations). The key conclusion that arises from the analysis of these baselines is that word lexemes are more informative than higher-level features such as POS-tags. Another insight concerns the high degree of redundancy present in the corpus, specifically in the consistent use of the same labels (i.e, Action or Concept) for the same words across the train, development and test sets. For this reason, it is easier to obtain a larger precision than recall. More details about the baseline implementations and their results are provided in the Appendix section.

The participating systems displayed a wide variety of approaches, such as classic supervised learning, deep learning, specialized knowledge bases and handcrafted rules. Almost all systems applied classic natural language processing techniques as a pre-processing step. For the identification of key phrases and their classification, the most interesting approach is based on a joint phrase recognition and classification using a BI-LSTM as a feature extractor and a CRF model for final classification [25]. For the relation extraction, the most promising approaches are based on convolutional neural networks, using a variety of features from morphological and syntactic to word vectors [26, 27]. The best performing system obtained a score of $F_1 = 0.646$ in the overall evaluation, while the best individual results per task were $F_A = 0.872$, $F_B = 0.959$ and $F_C = 0.448$. These results show that the relation extraction subtask is a challenging problem, where achieving significant progress will likely require more advanced machine learning techniques.

6. Discussion

The eHealth-KD corpus presents SAT+R, a general annotation structure, applied to documents from the health domain. This approach is not common, since health-related corpora are mostly annotated with domain-specific semantics. However, it is a useful approach for extracting relevant semantic knowledge in this domain, even if the relations and entities defined are not particular to the domain. Furthermore, the fact that the semantic annotations are general allows for directly applying the same schema to other sources in different domains, while being able to reuse all the learning algorithms, evaluation metrics, or in general, software based on this schema.

To allow for a broad range of semantics, 4 types of relations were defined in Section 3. They encode common semantic relations in general purpose knowledge bases, such as *hyperonyms* (**is-a**), *meronyms* (**part-of** and **property-of**) and *holonyms* (**same-as**). Their generality makes these relations likely to be relevant in most, if not all, knowledge domains. However, these relations alone are not enough to capture most semantics, as shown by the large number of **Action** tags present in the corpus. The Subject-Action-Target structure is general enough to capture a large part of a document's semantics without requiring a domain-specific conceptualization.

A significant sample of the semantics were correctly captured by the proposed schema. However, during the tagging process, annotators were able to identify several frequently occurring semantic patterns which were not captured by the semantic structure defined. Two of the most recurring patterns involve temporal, spatial and causal relations. This hints at the possibility of including categories of **Time** and **Location** and the corresponding relations of **occurs-at** and **located-in** for relating Concepts and/or Actions with these new semantic elements. Likewise, semantic relations such as **how**, **through**, and **entails**, to cover the possible causal connections between different Actions. Besides these specialized Action-Action relations, we also plan to annotate the previously described semantic relations between Actions (**is-a**, etc.) where necessary. These additions would allow the proposed semantic structure to cover a larger part of the semantics of a broad range of knowledge domains, without the need to resort to domain-specific categories or relations.

The largest difficulty during the annotation process consisted of correcting syntactic and semantic errors. In the initial phases of the process, many annotators frequently attempted to tag by unconsciously following a syntactic heuristic: i.e., the verb of the sentence is selected as Action, and the subject and direct object as Subject and Target respectively. Although this heuristic can work, in many cases it leads to incorrectly annotating as Subject-Action-Target something that

is best described by one of the semantic relations (is-a, etc.). In other cases, the sentence was written in passive voice or had a complex syntactic structure. For these reasons, annotators were carefully supervised during the initial period, and frequently corrected, until a common ground was reached.

The final version of the corpus is largely consistent and has been thoroughly revised by several annotators. Each final sentence has been viewed by at least 5 different annotators and was incorporated provided that agreement was reach among at least three of them (Stage 4). Hence, even though the defined F_1 metric provides an adequate quantitative measure of inter-annotator agreement, in a qualitative analysis it can be argued that the corpus quality is high. Even so, the use of the corpus by participants in the TASS 2018 Shared Task 3 provided useful insights for improving its consistency. Some participants detected patterns of annotation that can be improved, mostly related with inconsistencies between the subject and target roles in sentences with a complex structure. As described previously, these were among the most complicated situations for reaching agreement among annotators. In future versions of the corpus we will ensure that more clearly defined instructions and examples are provided to reach a more consistent annotation. Another reason that complicates learning is the relative unbalance of the different relation types. Since the corpus sentences were uniformly sampled from Medline, rare relations such as same-as appear much less than target, subject and is-a instances. Likewise, even though the train and test collections were sampled to present a 50% overlap, no explicit effort was taken to guarantee that the relative ratio of classes was the same or similar.

With respect to the complexity of automatically learning the corpus semantics, the baselines implemented show that pure syntactic and morphological features are not enough to achieve a high performance. The competition results suggest that some combination of deep learning techniques with pre-trained word embeddings, and a careful selection of additional morphological and syntactic features provide a much higher performance. On the other hand, knowledge-based approaches which are aided by external knowledge bases also perform competitively. According to the participants, the most complex patterns to learn are associated with the extraction of overlapping concepts (and the corresponding is-a relations), since several of the presented approaches could not deal directly with overlapping key phrases. This opens the door for trying a combination of statistical machine learning coupled with domain-specific knowledge, perhaps in the form of pre-trained embeddings learned from text corpora of a similar domain.

7. Conclusions and Future Work

This paper presents a corpus of Spanish health-related sentences, annotated with a generic Subject-Action-Target conceptualization. The corpus was tagged by various annotators, using an iterative process designed to maximize the consistency of the annotations and eliminate the most ambiguous sentences. The annotation is based on the semantic interpretation of the sentences, using an Action-Subject-Target structure with additional semantic relations. Several quality metrics were evaluated, demonstrating that the corpus is a reliable tool for training knowledge discovery systems in the Health domain. Considering the complexity of learning the corpus semantics, three different baseline algorithms were deployed, which exploit different characteristics of the sentences. Furthermore, the corpus was used for the evaluation of a shared task, where participants presented a variety of learning techniques, showing promising results.

During the annotation phase, we identified possible modifications to the annotation model. These will be evaluated and we will consider applying them in the next steps of this research,

which may result in the addition of more annotated sentences. A further addition to the annotation schema is the use of co-references to link the same Concepts and Actions across sentences. The annotation structure defined is not limited to the health domain, and therefore, it would be interesting to apply it to other text corpora in a different knowledge domain. We will also explore the process of linking the annotated Concepts and Actions with specialized knowledge bases, such as DBPedia or UMLS. These tasks aim to create useful resources in the field of automatic knowledge acquisition.

Conflict of Interests

The authors report there are no conflict of interest.

Acknowledgments

Funding: This research has been supported by the University of Alicante and University of Havana. Moreover, it has also been partially funded by both aforementioned universities and the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the projects PROMETEO/2018/089, PROMETEU/2018/089; Social-Univ 2.0 (ENCARGO-INTERNOOMNI-1); and PINGVALUE3-18Y.

This version of the paper takes into account helpful comments provided by the anonymous reviewers.

References

[1] E. Martínez-Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, V.-R. Julio, Overview of TASS 2018: Opinions, health and emotions, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018), volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS, Sevilla, Spain, 2018.

[2] P. Cimiano, A. Mädche, S. Staab, J. Völker, Ontology learning, in: Handbook on ontologies, Springer, 2009, pp. 245–267.

[3] K. Barker, B. Agashe, S. Y. Chaw, J. Fan, N. Friedland, M. Glass, J. Hobbs, E. Hovy, D. Israel, D. S. Kim, et al., Learning by reading: A prototype system, performance baseline and lessons learned, in: AAAI, volume 7, pp. 280–286.

[4] G. Gonzalez-Hernandez, A. Sarker, K. O'Connor, G. Savova, Capturing the patient's perspective: a review of advances in natural language processing of health-related text, Yearbook of medical informatics 26 (2017) 214–227.

[5] S. Estevez-Velarde, Y. Gutierrez, A. Montoyo, A. Piad-Morffis, R. Munoz, Y. Almeida-Cruz, Gathering object interactions as semantic knowledge (accepted), in: Proceedings of the 2017 International Conference on Artificial Intelligence (ICAI'17).

[6] A. Balahur, J. M. Hermida, A. Montoyo, R. Muñoz, Emotinet: A knowledge base for emotion detection in text built on the appraisal theories, in: R. Muñoz, A. Montoyo, E. Métais (Eds.), Natural Language Processing and Information Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 27–39.

[7] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, et al., Never-ending learning, Communications of the ACM 61 (2018) 103–115.

[8] T. C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, Journal of biomedical informatics 36 (2003) 462–477.

[9] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[10] F. Giunchiglia, M. Fumagalli, Teleologies: Objects, actions and functions, in: H. C. Mayr, G. Guizzardi, H. Ma, O. Pastor (Eds.), Conceptual Modeling, Springer International Publishing, Cham, 2017, pp. 520–534.

[11] G. Miller, WordNet: An electronic lexical database, MIT press, 1998.

[12] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web 6 (2015) 167–195.

[13] R. Speer, C. Havasi, Representing general relational knowledge in conceptnet 5., in: LREC, pp. 3679–3686.

[14] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract meaning representation for sembanking, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186.

[15] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, G. Zuccon, Overview of the clef ehealth evaluation lab 2016, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, pp. 255–266.

[16] M. Fabian, K. Gjergji, W. Gerhard, et al., Yago: A core of semantic knowledge unifying wordnet and wikipedia, in: 16th International World Wide Web Conference, WWW, pp. 697–706.

[17] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions, Journal of biomedical informatics 46 (2013) 914–920.

[18] I. Moreno, E. Boldrini, P. Moreda, M. T. Romá-Ferri, Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics, Journal of biomedical informatics 72 (2017) 8–22.

[19] A. Intxaurrondo, J. de la Torre, H. Rodriguez Betanco, M. Marimon, J. Lopez-Martin, A. Gonzalez-Agirre, J. Santamarıa, M. Villegas, M. Krallinger, Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of spanish clinical abbreviations: the barr2 corpus, SEPLN.

[20] M. Oronoz, K. Gojenola, A. Pérez, A. D. de Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions, Journal of biomedical informatics 56 (2015) 318–332.

[21] J. May, J. Priyadarshi, Semeval-2017 task 9: Abstract meaning representation parsing and generation, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 536–545.

[22] I. Augenstein, M. Das, S. Riedel, L. Vikraman, A. McCallum, Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 546–555.

[23] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 102–107.

[24] J. L. Fleiss, J. Cohen, B. Everitt, Large sample standard errors of kappa and weighted kappa., Psychological bulletin 72 (1969) 323.

[25] R. M. R. Zavala, P. Martínez, I. Segura-Bedmar, A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018), volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS, Sevilla, Spain, 2018.

[26] S. Medina, J. Turmo, Joint classification of key-phrases and relations in electronic health documents, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018), volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS, Sevilla, Spain, 2018.

[27] V. Suarez-Paniagua, I. Segura-Bedmar, P. Martínez, Labda at tass-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018), volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS, Sevilla, Spain, 2018.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[29] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear (2017).

Appendix

In this appendix we present more details about the evaluation metrics used in the TASS 2018 eHealth-KD challenge [1] and the baseline implementations.

Evaluation metrics

To compute the evaluation metrics for each subtask, the following sets were defined, for the annotations between the gold output and the actual output in each subtask:

Correct matches (C): in all tasks, when one gold and one given annotation match exactly.

Partial matches (P): in Task A, when two key phrases have a non-empty intersection.

Missing matches (M): in Task A and C, when an annotation in the gold output is not provided by the system.

Spurious matches (S): in Task A and C, when an annotation given by the system does not appear in the gold output.

Incorrect matches (I): in Task B, when one assigned label is incorrect.

Given these criteria, we define overall precision and recall as follows:

$$precision = \frac{C_A + \frac{1}{2}P_A + C_B + C_C}{C_A + M_A + P_A + C_B + I_B + C_C + M_C}$$

$$recall = \frac{C_A + \frac{1}{2}P_A + C_B + C_C}{C_A + S_A + P_A + C_B + I_B + C_C + S_C}$$

The final evaluation metric is a standard F_1 measure defined as:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Similarly, individual F_1 metrics for each subtask can be calculated to evaluate the quality of a particular phase in a given system. The test collection provides separate output files for each subtask, which allows researchers to reuse the gold outputs of, say, subtasks A and B, and focus on improving performance on subtask C.

Baseline implementations

In order to define a baseline comparison for other learning systems trained on the corpus, three basic strategies are evaluated. These strategies include a dummy approach based solely on the text of key phrases, a statistical approach based on word vectors and a decision tree-based approach using syntactic and semantic features[9]. The implementations are based on `scikit-learn` [28]. Further details are provided below.

[9] https://github.com/knowledge-learning/ehealth-kd/tree/master/baseline

	Development	Test			
Metric	**ABC**	**A**	**B**	**C**	**ABC**
B1- Precision	0.708	0.673	0.774	0.714	0.755
B1- Recall	0.331	0.536	-	0.058	0.452
B1- F1	0.452	0.597	-	0.107	0.566
B2- Precision	0.485	0.808	0.899	0.175	0.496
B2- Recall	0.625	0.756	-	0.663	0.666
B2- F1	0.546	0.781	-	0.277	0.568
B3- Precision	0.195	0.794	0.936	0.106	0.224
B3- Recall	0.733	0.824	-	0.870	0.744
B3- F1	0.308	0.809	-	0.189	0.345

Table 5: Results of the baseline algorithm.

B1 - Dummy: This technique collects all training data and stores three maps: (1) from key phrases to the most common class (either Concept or Action); (2) pairs of concepts to their most common relation; and (3) tuples of <Action,Concept> to their most common role. At prediction time, these maps are used to select a key phrase, decide its class, and predict relations and roles.

B2 - Word vectors: In this technique each key phrase is represented by its standard vector embedding (using Spanish Word2Vec values from `spaCy` [29]). The word vector of a multi-word phrase is computed as the average of the word vectors of its components. Three logistic regression models are trained with this representation: (1) a binary classifier that decides if the key phrase is relevant or not (for task A and B); (2) a classifier that predicts the relations between a pair of word vectors; and (3) a classifier that predicts the corresponding role for a pair of vectors.

B3 - Decision Tree: In this technique each key phrase is represented by a collection of its syntactic and semantic characteristics, including part-of-speech labels, genre, person, dependency label in the dependency tree of the sentence, among other linguistic features. The actual text of the key phrases is *not* included. With this representation, three classifiers are trained, as in baseline **B2**, only this time standard decision trees are used.

Table 5 summarizes the results of the three baseline implementations. All implementations are trained on the training collection only, and then tested on the development and test collections. Individual task results for B and C are obtained by using the gold output of the previous task (A or B) as inputs. The general trend shows that Task B is the easiest, followed by Task A and then Task C. This is to be expected, since Task B is a classic binary classification problem for which standard techniques suffice. Task A is at least adequately solved, hence, it is in Task C where most of the innovative research is likely to happen.

The baseline dummy implementation (B1) obtains a larger precision compared to recall. This is an indication that the same key phrases are mostly used consistently in the corpus. Hence, when entities present in the training collection appear in the test collection, the corresponding relations are mostly the same. However, there are many pairs of entities in the development and test collections that do not appear in the training collection. To solve these pairs, we expect that additional semantic features must be exploited that do not rely solely on the key phrases lexemes.

The two baseline implementations based on machine learning, (B2 and B3), obtain a larger recall compared to precision. This is an indication that both models are producing a large number of false positives. By design, both models do not include the actual lexeme of the phrases but focus on higher-level features. However, embeddings do include some indirect representation of the word lexeme in the weights of the embedded vector, and this is evident in the precision variations between both models when applied to Task A. In conclusion, the higher-level features (syntactic and semantic) were found to be less relevant than lexemes.