This is a previous version of the article published in Journal of Biomedical Informatics. 2020, 109: 103517. https://doi.org/10.1016/j.jbi.2020.103517

A Computational Ecosystem to Support eHealth Knowledge Discovery Technologies in Spanish

Alejandro Piad-Morffis[a,*], Yoan Gutiérrez[b,c], Yudivian Almeida-Cruz[a], Rafael Muñoz[b,c]

[a] *Faculty of Math and Computer Science, University of Habana, La Habana, Cuba, 10200.*
[b] *University Institute for Computing Research (IUII), University of Alicante, Alicante, Spain, 03690.*
[c] *Department of Language and Computing Systems, University of Alicante, Alicante, Spain, 03690.*

Abstract

The massive amount of biomedical information published online requires the development of automatic knowledge discovery technologies to effectively make use of this available content. To foster and support this, the research community creates linguistic resources, such as annotated corpora, and designs shared evaluation campaigns and academic competitive challenges. This work describes an ecosystem that facilitates research and development in knowledge discovery in the biomedical domain, specifically in Spanish language. To this end, several resources are developed and shared with the research community, including a novel semantic annotation model, an annotated corpus of 1,045 sentences, and computational resources to build and evaluate automatic knowledge discovery techniques. Furthermore, a research task is defined with objective evaluation criteria, and an online evaluation environment is setup and maintained, enabling researchers interested in this task to obtain immediate feedback and compare their results with the state-of-the-art. As a case study, we analyze the results of a competitive challenge based on these resources and provide guidelines for future research. The constructed ecosystem provides an effective learning and evaluation environment to encourage research in knowledge discovery in Spanish biomedical documents.

Keywords: Knowledge Discovery, Annotated Corpora, Semantic Annotation Models, Entity Recognition, Relation Extraction, Natural Language Processing

1. Introduction

The exponential growth of the Internet in the last decades has produced a massive surplus of textual information in all areas of human endeavor. This scenario presents both an opportunity and a challenge for researchers. On the one hand, a growing body of scientific literature is readily available, where potential solutions for critical problems could be found by linking partial results published in distinct documents. On the other hand, the extent of the information available cannot be processed by humans alone in a reasonable time frame. Hence, efforts have recently been directed towards designing automatic techniques that can discover relevant pieces of information

*Corresponding author.
Email addresses: apiad@matcom.uh.cu (Alejandro Piad-Morffis), ygutierrez@dlsi.ua.es (Yoan Gutiérrez), yudy@matcom.uh.cu (Yudivian Almeida-Cruz), rafael@dlsi.ua.es (Rafael Muñoz)

in large corpora, make logical connections, and synthesize useful knowledge. The first step in many of these techniques involves the collection, processing and annotation of data that can be used to train machine learning algorithms or build expert systems through the use of natural language processing techniques.

The digital health sector is of great interest to the research community given the potential social benefits derived from applying automatic knowledge discovery technologies. The research community has produced a large number of annotated corpora in different sub-domains of this sector, from specific (e.g., drug-disease [1] or gene-protein interactions [2]) to broad in scope and domain (e.g., clinical trial reports [3]). Domain-specific corpora and technologies are of critical importance in high-precision medicine. However, systems built for very specific domains are arguably harder to generalize and extend than systems built on general-purpose conceptualizations. As such, there is a growing interest in designing annotation models and corpora with general-purpose semantics that can be used in a variety of domains or as a component in more specialized systems.

Besides domain, language is another dimension that has been the focus of recent research. Most of the largest linguistic resources are based on English sources, motivated in part by the abundance of available raw material (e.g., online encyclopedias, research papers), which is not surprising given that English is the most predominant language in science, technology and communications. However, English-based resources are not always directly applicable to other languages. Even though automatic translation has reached impressive accuracy in open domains, it is still a challenge to create cross-language resources, such as with Spanish, which is less predominant in technical domains [4]. Instead of focusing on specific niche languages, one possible line of research is designing resources that are language-agnostic, in the sense that they can be generalized to multiple languages with little effort, by virtue of being based on underlying common characteristics shared by many languages.

Designing annotation models that can generalize to multiple domains requires deciding on a basic representation of language that covers a broad range of semantics. Moreover, these representations should be as independent of syntax and grammatical rules as possible, if they are expected to generalize to multiple languages. Recent work [5] suggests that Subject-Action-Target triplets can be used to detect a large number of semantic interactions in natural language, independent of domain and relatively independent of language, since more than 75% of human languages employ some variation of the Subject-Verb-Object grammatical structure [6]. Likewise, several ontological representations often agree in a number of general-purpose relations, (e.g., *is-a* hyponyms, *part-of* holonyms) that are useful in any domain. Other conceptualizations allow the capture of semantics closer to natural language, such as Abstract Meaning Representation, AMR [7]. The construction of corpora annotated with general-purpose semantic structures like Subject-Action-Target and high-level ontological relations is the first step in the design of systems that can discover knowledge automatically in a variety of domains and scenarios.

Research in knowledge discovery requires not only linguistic resources (e.g., annotated corpora) but also computational resources and infrastructures that enable researchers to systematically evaluate their results and compare them objectively with alternative approaches. This involves the formal definition of tasks and the design of objective evaluation metrics that ensure fair comparison is possible. Even better is a publicly available evaluation system where researchers can submit their results, guaranteeing the same evaluation criteria is applied and freeing researchers from reproducing the evaluation environment. Such a system would also guarantee a more transparent and reproducible research process, and would provide a centralized repository of existing approaches, helping new researchers to update on the state-of-the-art.

This research focuses on the construction of an ecosystem for supporting the development of eHealth Knowledge Discovery (eHealth-KD) technologies. This ecosystem consists of linguistic resources, such as the definition of an annotation semantic model and corpora; tools and infrastructure for deploying and testing systems; and, evaluation metrics to allow fair comparisons. Concretely, the contributions of this research are:

- The definition of a semantic model and a related annotation model to capture broad sentence semantics in natural language text.
- The development of the *eHealth-KD v2* corpus [8], a manually annotated corpus of Spanish language sentences in the health domain, and an analysis of its characteristics and quality metrics.
- A formal definition of a knowledge discovery task based on this corpus, as well as evaluation metrics for two different subtasks of interest.
- The development of an infrastructure to support the creation of systems for the aforementioned task, including baseline systems and tools; and an online service for the automatic and continuous evaluation of new techniques.
- An in-depth analysis of several existing systems evaluated in this ecosystem, providing insights on the most promising strategies and outlining interesting directions for future research.

The remainder of the paper is organized as follows. Section 2 presents the most relevant related works in the scientific literature, including annotated corpora, technologies and tools to support the development of knowledge discovery systems, and evaluation scenarios, campaigns and challenges in this area. Section 3 introduces the annotation model used in the *eHealth-KD v2* corpus, the annotation process and the main characteristics of the corpus obtained. Section 4 defines a computational task based on the corpus with objective evaluation metrics, and describes an existing infrastructure available for researchers aiming at solving the proposed task. Section 5 analyses existing systems for solving this task, highlighting the most promising approaches. Section 6 discusses the most relevant aspects of the whole research, lessons learned, and limitations. Finally, Section 7 presents the conclusions and recommendations for future research.

2. Related Work

In this section we analyze the elements that contribute to a successful research in knowledge discovery, specifically in health-related domains. Section 2.1 discusses relevant linguistic resources available for researchers in this area, including annotated corpora and related semantic models, both in general-purpose domains and specifically for the health domain. Section 2.2 presents a brief comparison of existing technologies to support the construction of linguistic resources, i.e., annotation tools. Finally, Section 2.3 explores the role of competitive evaluation campaigns and challenges in fostering research in this area, and summarizes previous efforts in this respect.

2.1. Linguistic Resources for Knowledge Discovery in eHealth

Different semantic relations have been established in the state of the art, many of these giving rise to the construction of corpora. We focus on two approaches: corpora or annotation models to represent knowledge in many domains as well as those specifically about health. The table 1 presents the seven characteristics relevant to our corpus and indicates which of them are present in a sample of corpora from the state-of-the-art. These characteristics can be understood in the following terms:

1. *general-purpose annotation:* applicability of the underlying annotation model to any domain;
2. *independence of syntax:* capturing semantic aspects rather than syntactic relations in sentences;
3. *ontological knowledge:* supporting inheritance and composition between concepts;
4. *composite concepts:* allowing the annotation of concepts that involve other sub-concepts;
5. *attributes:* modeling attributes for each annotated entity such as quantifiers (e.g., number of occurrences) or qualifiers (e.g., degree of certainty);
6. *contextual relations:* modeling relations that only occur when conditioned by a specific context; and,
7. *causality / entailment:* including relations for representing causality and/or entailment.

	Characteristics	**Ixa MedGS** [9]	**DrugSemantics** [10]	**DDI** [11]	**Bio AMR** [12]	**YAGO** [13]	**ConceptNet** [14]	**eHealth-KD v1** [15]	**eHealth-KD v2**
1	general-purpose annotation				✓	✓	✓	✓	✓
2	independence of syntax	✓	✓	✓		✓	✓	✓	✓
3	ontological knowledge				✓	✓	✓	✓	✓
4	composite concepts				✓			✓	✓
5	attributes		✓		✓	✓		✓	✓
6	contextual relations				✓				✓
7	causality / entailment	✓			✓		✓		✓

Table 1: Comparison between the *eHealth-KD v2* corpus and other corpora with respect to the characteristics that define our proposal.

General-purpose annotation. General-purpose annotation models are often used in corpora extracted from encyclopedic sources, such as *YAGO* [13] and *ConceptNet* [14], both of which contain facts automatically extracted from Wikipedia (among other sources). In contrast, domain-specific annotation models are usually employed when the source is more restricted to a specific domain. Examples include *Ixa MedGS* [9], which contains health related concepts for diseases, causes and medications; *DrugSemantics* [10], which annotates health entities, drugs and procedures; and, *DDI* [11], which annotates drug-drug interactions. A middle ground is the *Bio AMR* [12] corpus, which applies a general purpose annotation model (AMR) [7] to health documents. The *eHealth-KD v2* corpus is similar to the latter in this respect, since the annotation model defined is general, but it is applied specifically to health sentences in this research. The *eHealth-KD v2* corpus constitutes the result of the evolution of the *eHealth-KD v1* [15] corpus.

Most of the aforementioned resources are focused on capturing the semantics of sentences, in the sense that very different sentences with the same facts are likely to be similarly annotated. We consider *BioAMR* less independent of syntax because even though AMR is a semantic annotation model—far more abstract than dependency parsing, for example—, it still relies heavily on sentence grammatical structure. Hence, a significant change in the sentence structure is likely to change the annotation, even if the underlying semantic message remains unchanged. For example, since AMR uses PropBank [16] roles, changing a word for a semantically similar word, including a synonym, will probably change the corresponding annotation and thereby the available roles. This also makes AMR and similar resources language-dependent, not only in practice given their dependence on the existence of word banks, but also in nature. While attempting to apply AMR in Spanish, Migueles-Abraira et al. [17] show that even though AMR is theoretically language-agnostic, the existing annotation guidelines are biased towards English and must be adapted to capture linguistic phenomena that don't exist in English. The annotation model designed in this research for the *eHealth-KD v2* corpus, attempts to achieve a higher level of syntactic independence, in part by using a smaller set of entities, relations and roles than AMR. More specifically, our annotation model does not distinguish semantic roles for each possible `Action`, instead relying on general purpose roles (i.e., `subject` and `target`, see Section 3.1).

Ontological knowledge. General-purpose annotation models often allow ontological knowledge to be represented in the form of inheritance and composition between concepts. In this context, we consider the ability to recognize and annotate these ontological relations in the source text. Health-related annotation models do not usually deal with this problem, mainly because the entities and relations to annotate form a predefined ontology where composition and hierarchy, if any exist, are already conceived in the annotation model itself. However, general purpose annotations often include relations like *ConceptNet*'s `is-a` or `part-of` that directly represent these ontological concepts, and are thus able to extract ontological representations from natural text.

Composite concepts. The model designed for the *eHealth-KD v2* corpus also includes relations specifically for this purpose, mostly inspired by *ConceptNet* and *YAGO*. Composite concepts, in contrast, refer to the ability to annotate concepts that are formed by a fine-grained combination of other entities, in the same sentence. For example, take the sentence: "*the doctors that work the night shift get paid extra hours*". *AMR* allows for the representation of the concept that not all doctors, but only those that work the night shift, are the ones who get paid extra hours. Our proposal also includes several annotation patterns to deal with this type of scenario.

Attributes. Attributes are often used to further refine the meaning of annotated entities. Examples include quantifiers in *AMR*, or modifiers that specify a degree of uncertainty, or a negation of a concept. Our proposal includes four general-purpose attributes that model uncertainty, negation and qualifiers for expressing emphasis.

Contextual relations. Contextual relations, as defined in the *eHealth-KD v2* corpus, allow facts that only occur under certain conditions to be represented, for example, in a specific time frame or location or under certain assumptions. This allows for a finer-grained semantic annotation. *BioAMR* inherits this ability from *AMR*, which allows modifiers for expressing *how*, *when*, *where* or *why* some event occurs. In our proposal, we provide contextual relations that specify time and location, and an additional general-purpose relation for other conditions.

Causality and entailment. Causality and entailment are general-purpose relations that allow some level of inference or reasoning. The *Ixa MedGS* corpus defines a `causes` relation, since it is relevant in the domain the corpus is modeling. Likewise, *AMR* and *ConceptNet* include similar relations. Our proposal includes both causality and entailment as two different relations with well-defined semantic meanings.

2.2. Technologies for Annotation and Resource Distribution

An important element to consider in Knowledge Discovery research is the existence of computational resources and infrastructure that supports the development of new approaches. The creation of linguistic resources often stems from a process of manual annotation by human experts, which requires computational tools for the actual annotation as well as mechanisms for merging annotations and computing agreement, ideally in a collaborative environment. Once the resources are created, it is necessary to distribute the corresponding corpus, baselines, and tools among the research community, often through online source code sharing platforms.

An extensive analysis and comparison of several annotation tools is provided in Neves and Ŝeva [18]. Table 2 summarizes the main characteristics we considered relevant for this research and identifies the most appropriate annotation tool among a subset of popular alternatives. We consider as requisites web-based, open source annotation tools that allow multi-label span annotations as well as relation annotations. Support for collaborative annotation, at least partially, is also highly desirable. Of the analyzed tools, we identified Brat [19] and WebAnno [20], as they comply with all the aforementioned requisites. In our research, we preferred Brat to WebAnno because, even though WebAnno provides more features, Brat allows an easier setup. It is not only faster to start an annotation project using this tool, but also to train annotators to use its interface.

The public distribution of annotated corpora and related resources, e.g., baselines, evaluation scripts, loading and formatting scripts, etc., is often enabled via open source code sharing platforms. Arguably the most popular options are Github[1] and Gitlab[2], which provide similar features despite minor differences in their core business models. It is also possible to share the corresponding resources via institutional hosting platforms or other ad-hoc solutions. This could be convenient in the case of legal requirements, complex licenses that are incompatible with open source idiosyncrasies or any other consideration that disallows full public sharing. In our case, all resources are publicly available in a collection of Gitlab repositories[3].

[1] https://github.com
[2] https://gitlab.com
[3] https://ehealthkd.gitlab.io

Characteristics	GATE Teamware	Knowtator	WebAnno	Brat	BioQRator	CATMA	prodigy	TextAE	LightTag	Djangology	MyMiner	WAT-SL
multi-label annotations			✓	✓		✓			✓	✓		
relation annotations		✓	✓	✓	✓			✓	✓		≈	
allows custom model	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
collaborative interface	✓		≈	≈	≈	≈	≈		✓	✓		≈
web-based interface	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
can be self-hosted	✓	✓	✓	✓		✓	✓	✓		✓		✓
open source license	✓	✓	✓	✓		✓		✓		✓	✓	✓
citation	[21]	[22]	[20]	[19]	[23]	[24]	[25]	[26]	[27]	[28]	[29]	[30]

Table 2: Qualitative comparison of popular annotation tools. Adapted from Table 3 in Neves and Ŝeva [18], Table 3. A symbol ≈ indicates that the corresponding feature is only partially supported.

2.3. Evaluation Campaigns

A strategy often used to encourage research on a specific task is the organization of a shared evaluation campaign. In contrast with regular research, evaluation campaigns often have a fixed time frame, and evaluation resources are not fully disclosed (e.g., gold annotations for test sets are hidden) to allow a fair comparison in a friendly competitive environment. In this section, we analyze relevant efforts for organizing evaluation campaigns for both the biomedical domain or for dealing with entity and relation extraction.

Several online services allow researchers to organize machine learning challenges and competitions, providing automatic grading, user management, and other useful features. Kaggle[4] is arguably the most popular choice, its main limitation for our purposes being that to host a challenge, organizers must contact the service providers. Possible alternatives are AIcrowd[5] and Codalab[6] which provide free options for challenge organizers.

The CLEF eHealth Evaluation Lab has proposed several challenges in the biomedical domain, including named entity recognition [31] and information extraction [32] in English, and later editions in French documents [33, 34]. In these challenges, medical reports from MEDLINE, EMEA and similar sources are annotated with disorders, medical terms, acronyms and abbreviations, which provide evaluation scenarios for several NLP tasks, including entity recognition, normalization and disambiguation. Another relevant task is proposed by May and Priyadarshi [35] in Semeval 2017, focused on AMR parsing and generation from biomedical sentences in English. Applying a general-purpose conceptualization, such as AMR, to specific domains encouraged participants to bridge the gap between developing generalizable techniques and applying domain-specific heuristics. However, AMR parsing is already a complex problem in itself, which can negatively impact on researcher participation in these challenges if they are not specialized in AMR. Simpler, general-purpose models can encourage a greater degree of participation given the easier entry curve. An example of the latter is the Semeval 2017 Task 10 [36], a challenge regarding keyphrase and relation extraction from scientific documents, with a simple model based

[4] https://kaggle.com
[5] https://www.aicrowd.com
[6] https://codalab.org

on three entity classes and two general-purpose relations. This task received a much larger number of submissions than the former, even though both challenges where hosted on the same venue and aimed at similar audiences.

As can be expected, English is the most prominently used language in NER-related challenges, given the larger number of available corpora and resources. However, important efforts have been devoted to fostering research in less prominent languages. Relevant to our discussion are the IberLEF campaigns that focus on Iberian languages, such as Spanish, Portuguese, Catalan, and other regional variations. Two examples of recent NER-related tasks are the Portuguese Named Entity [37] challenge and the MEDDOCAN [38] document anonymization challenge. The first proposes entity recognition and relation extraction in the general domain, in Portuguese. The second proposes the identification of privacy-sensitive entity mentions in medical documents, e.g., names, addresses, dates, ages, etc. Finally, related to the *eHealth-KD v1* and *v2* corpora, two challenges have been proposed, respectively in the TASS 2018 Workshop [39] and IberLEF 2019 [40] editions. These challenges introduced the task described in Section 4, which gathered significant attention from the NLP research community focused on processing Spanish language. Relevant results for the latest edition are discussed and analyzed in Section 5.

Outside the frame of a competition, open, long-running evaluation systems allow researchers to evaluate their approaches with official evaluation metrics. This can also provide a centralized repository of the state-of-the-art, where existing approaches are summarized and linked to existing papers. In this regard, this research proposes an online evaluation system that allows a comparison of new approaches with officially published results at any time. Based on this infrastructure, official evaluation campaigns with a more competitive design are organized in scheduled time-frames.

3. The eHealth-KD v2 Corpus

This section presents the *eHealth-KD v2* corpus, its main design decisions, annotation process, and relevant evaluation criteria. Section 3.1 describes a novel annotation model defined for this corpus that captures sentence-level semantics without resorting to domain-specific labels. Section 3.2 describes the annotation process of the corpus, and Section 3.3 presents a statistical analysis and relevant quality metrics. The corpus is available online for download[7] and shared in an open access repository [8].

3.1. Annotation Model

The annotation model defined for the *eHealth-KD v2* corpus draws inspiration from several resources. The main source of inspiration is the *eHealth-KD v1* corpus [15], annotated with a more restricted version of this model, whose main limitations in terms of expressibility are tackled by our proposal. In terms of knowledge representation, our annotation model draws from two different models for conceptualization of reality: ontologies and teleologies. For reference purposes, Figure 1 shows an example annotation of three sentences with various degrees of complexity. The annotation model is explained in-depth in Piad-Morffis et al. [41].

The ontological part of the model provides a representation of entities in the health domain in terms of hierarchical and structural relations (i.e., `is-a`, `part-of`, `has-property` and `same-as`). These relations are based on the design of upper ontologies such as ConceptNet [14]

[7] https://gitlab.com/ehealthkd/corpus

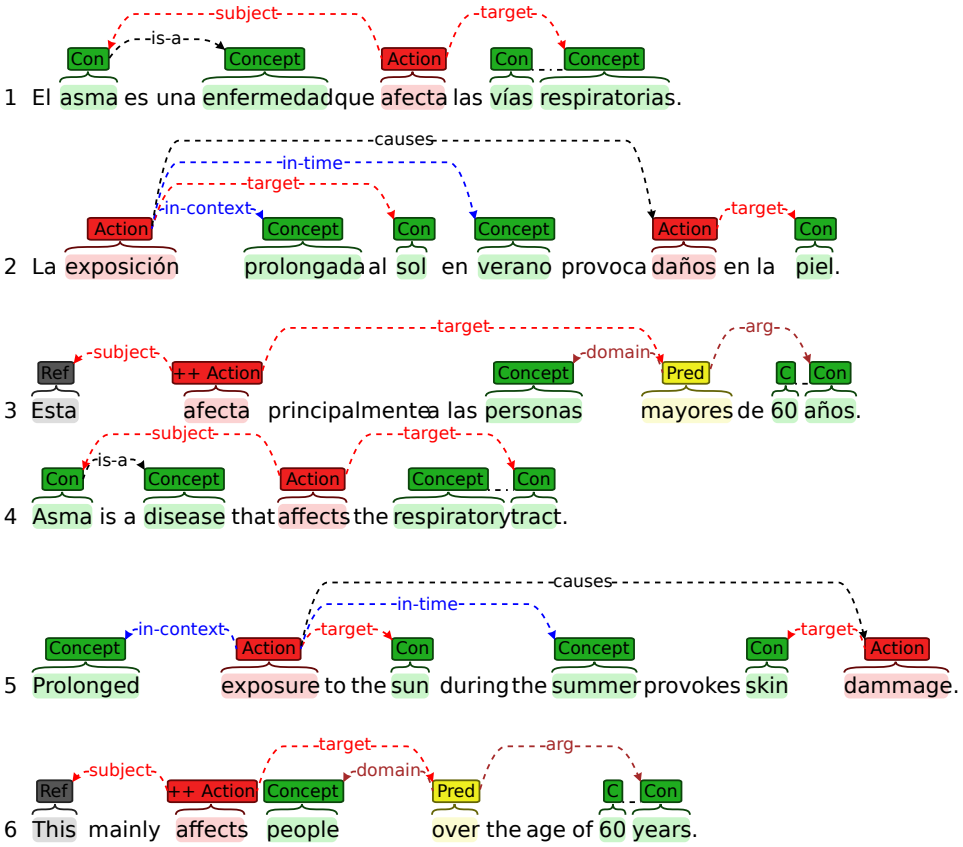

Figure 1: Example annotation of three sentences. The annotation shows the most relevant entities and relations defined. Adapted from Piad-Morffis et al. [41]. On the top, the original text in Spanish. On the bottom, for reference purposes, an English translation.

and YAGO Suchanek et al. [13]. The teleological part of the model provides a representation of events or processes in the health domain that transform entities, i.e., representing the purpose of things. This is supported by a Subject-Action-Target structure based on the work of Giunchiglia and Fumagalli [42]. The exact semantic meaning of these concepts and relations is further explained in this Section.

The core of the annotation is the structure Subject-Action-Target, which captures the main interaction in objective sentences. Two different entities participate in this interaction: `Concepts` and `Actions`. A `Concept` defines a relevant entity in the domain, which can either be a single word, or multiple tokens, contiguous or not. An `Action` represents a process or event caused by one or more `Concepts` (i.e., the `subjects`) and which impacts on one or more `Concepts` (i.e., the `targets`). The `subject` and `target` roles can also be `Actions` themselves, which enables simple concepts to be composed into more complex ones. The Subject-Action-Target structure defined in this model is based on a simplified version of the teleological framework by [42]. *Objects* and *Actions* in this framework are represented in our model by `Concept` and `Action` respectively. The *Function* role in teleologies, which expresses an instance of an object perform-

ing an action, can be approximately equated to our use of `Actions` as `subjects` or `targets` of other actions.

An important addition to this annotation model is the `Predicate` entity. Predicates model the existence of complex concepts (i.e., the `domain`) that are dependent on some preconditions (i.e., the `args`). For example, in Figure 1, Sentence 6, the concept of "*people over ... 60 years*"[8] can be defined with a fine-grained annotation, by considering "*people*"as the domain and "*60 years*" as the argument. This annotation allows the capture of more detailed information rather than simply annotating the whole phrase as a multi-word concept. Another addition is `References`, which represent unexplicitly mentioned concepts in a sentence. The most common words labeled as References are: "*esto*", "*el*", "*la*", " *este*", i.e., usually pronouns and articles.

To further refine the semantic interpretation of each entity, a set of 4 attributes is defined: `uncertain`, `emphasized`, `diminished` and `negated`. These attributes are often hinted at by adjectives or other syntactical patterns that appear outside the surface text of a given entity, but instead of annotating those phrases, the corresponding entity is tagged with the attribute. For example, in Sentence 6 of Figure 1, the action "*affects*" is attributed with `emphasized`, hinted by the world "*mainly*"[9] and represented in the annotation with a ++ sign in the action itself. The use of attributes allows the capture of more refined semantic concepts (i.e., degrees of emphasis, negation, uncertainty) while maintaining language-agnosticism, since it is irrelevant where in the surface text that information is presented. It can either be hinted explicitly by a single word (e.g., an adjective) or implicitly by a figure of speech, rhetorical language and other subtle linguistic cues. These attributes increase the range of semantics covered by the annotation model without increasing the number of tokens that need to be annotated.

In terms of relations, the *eHealth-KD v2* corpus inherits the 4 main ontological relations present in the previous version: `is-a`, `same-as`, `part-of` and `has-property`, with their usual semantics. Each of these relations can link any concept, both simple or complex with another. These relations allow the representation of structural knowledge, e.g., concepts related in a hierarchical structure, and concepts that are components of other concepts. Two additional relations are defined in this new version, `causes` and `entails`, to capture causality and logical entailment respectively. These relations, respectively teleological and ontological in nature, are of great importance because they enable the construction of reasoning systems that can reach conclusions and produce new knowledge from an existing corpora.

Additionally, 3 contextual relations are defined, to collect important knowledge that usually appears as a grammatical complement in sentences: `in-time`, `in-place` and `in-context`. The relation `in-time` is used for expressing the duration of an event. The relation `in-place` is used for identifying a specific location for the `Action` or `Concept`. The relation `in-context` is a more generic relation of this set, representing a general dependency on some other `Concepts` whose exact nature cannot defined by the annotation. These relations are also teleological in nature, as they do not define an assertion per-se, but instead are useful for specifying conditions in which some events occur. For example, in Sentence 5 (Figure 1, the annotation "*exposure* ⇒ `in-context` ⇒ *prolonged*"[10] does not imply that the concept "*exposure*" unconditionally has the quality "*prolonged*". It is only when this complex concept is used as `subject` or `target` of an `Action` or in another relation that the contextualization becomes meaningful.

[8] In Spanish: "*personas mayores de 60 años*", in Sentence 3.
[9] In Spanish, the corresponding word is "principalmente" in Sentence 3.
[10] In Spanish: "*exposición prolongada ...*" in Sentence 2.

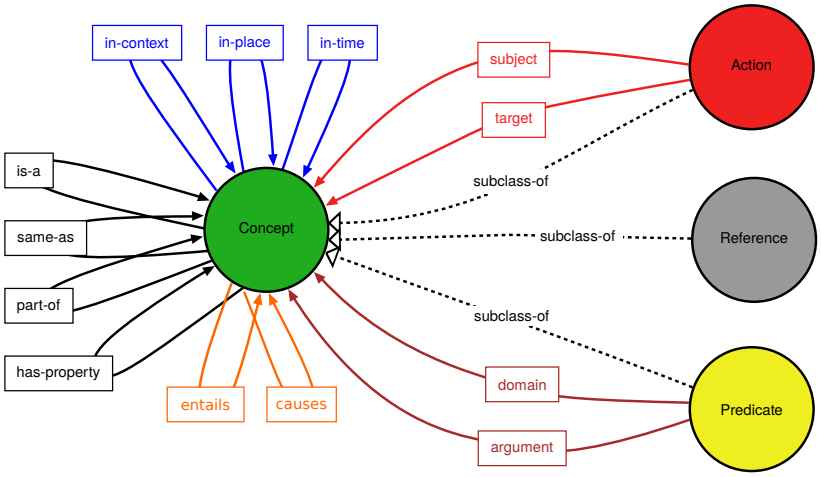

Figure 2: Conceptual schema for the annotation model. Each of the semantic roles defined in the annotation model are represented as circles. The possible relations defined between each pair of roles are represented as rectangles. Adapted from Piad-Morffis et al. [41]

.

Figure 2 summarizes the annotation model defined in the *eHealth-KD v2* corpus. This model is designed to be as general as possible to capture the most relevant semantic knowledge present in an arbitrary corpus. For this reason, no domain specific relations or entities were defined (i.e., no specific entities for diseases, patients, treatment, etc.). In contrast, domain specific relations can be represented via actions and their corresponding roles.

In comparison with the previous version of the annotation model, this new model extends its ability to annotate fine-grained concepts that are interrelated with each other. The previous version relied solely on `Action` and `Concept` for composition, and the 4 basic ontological relations. The addition of the `Predicate` annotation allows for a semantic differentiation between the main content of a sentence —what is being done by whom to who, indicated by `Actions`—, and additional descriptive content. The contextual relations provide additional fine-grained semantic meaning to common linguistic patterns. Furthermore, causality and entailment are completely new semantic relations that could not be expressed in the previous annotation model. A complete list of all new labels in this version is available in Table 3.

3.2. Annotation Process

The *eHealth-KD v2* corpus was built from a sample of Spanish language sentences taken from the MedlinePlus XML dumps[11]. The original source contains 2,026 entries in Spanish language of different topics related to health. Each entry was parsed, split into sentences, and filtered to remove unwanted content such as copyright notes, sentences ending in "?" and "!", sentences shorter than 5 words and HTML-specific content. Finally, a pool of 9,956 sentences was obtained, from which a random sample of 1,045 sentences was taken for the manual annotation process.

[11] `https://medlineplus.gov/xml.html`

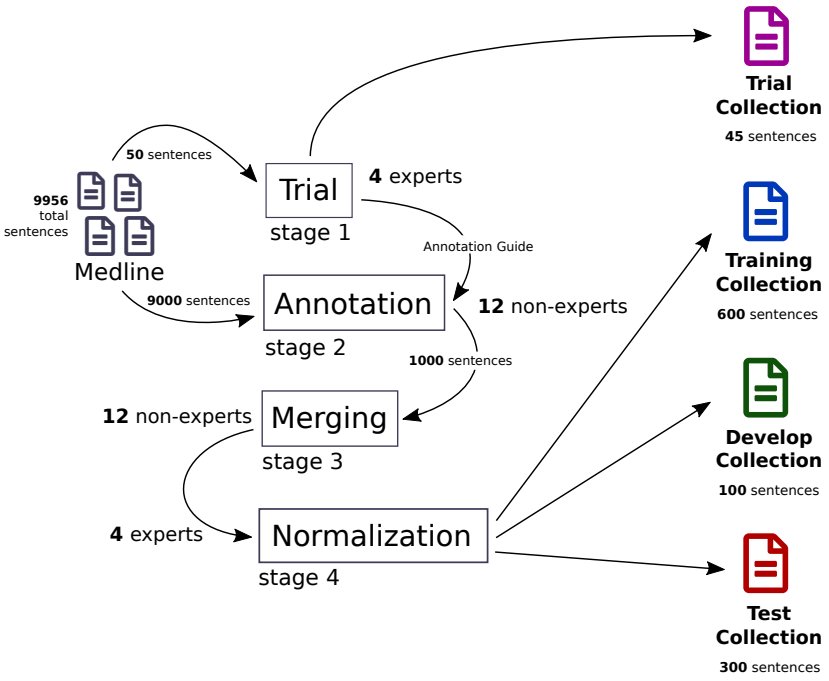

Figure 3: Schematic representation of the annotation process.

Before the main annotation, a small sample of 45 sentences was randomly selected and jointly annotated by a committee of 4 experts. This sample became the *trial* collection, and was used to produce suitable annotation guidelines and train a team of 12 non-expert annotators. The expert annotators are researchers and PhD students in the Natural Language Processing area, while non-experts are under-graduate students in Computer Science. By design, no domain-specific expertise or knowledge is required to correctly annotate the *eHealth-KD v2* corpus, since all semantics stem for the use of natural language and the source content for annotation is aimed at a general audience.

The rest of the corpus (i.e., 1,000 sentences) was manually annotated in 25 batches of 40 sentences, randomly selected from the previously described pool of sentences. Each batch was labeled by two different annotators independently. Each pair of annotations was subsequently merged automatically, with a custom tool developed for this purpose that fixes minor contradictions and highlights the remaining for a human annotator. A third independent annotator (one that didn't participate in that batch) was tasked with normalizing and fixing the remaining contradictions.

Afterwards, a committee of 4 experts reviewed all the sentences and for every case in which at least one of the reviewers did not agree with the annotation, the corresponding sentence was publicly discussed until agreement was reached. At this point, modifications to the annotation were allowed in rare cases and only if all reviewers agreed. The entire annotation process was performed in the Brat annotation tool [19] with the help of ad-hoc tools specifically built for the tasks of shuffling and filtering the sentences, merging the annotations, etc. Figure 3 shows a schematic representation of the complete annotation process.

The final version of the *eHealth-KD v2* corpus contains a total of 1,045 manually annotated and reviewed sentences. For the purpose of using the corpus in the development of knowledge discovery technologies, the sentences are divided into four collections: (1) *trial*, 45 sentences; (2) *training*, 600 sentences; (3) *validation*, 100 sentences; and (4) *test*, 300 sentences.

3.3. Corpus Statistics and Quality Metrics

Table 3 shows the number of entities and relations annotated in the *eHealthKD-2019* corpus. In total, 13,246 elements were annotated into 6,612 entities, 6,049 relations, and 585 attributes. The entities which appeared less were Predicate and Reference. The relations which appeared less were: `entails`, `in-time`, `has-property`, `same-as` and `part-of`.

For the entire corpus, the number of `Actions` is greater than the number of sentences, indicating that many sentences with more than one `Action` exist. In total, 222 complex concepts were annotated. These are `Action` annotations whose `target` or `subject` is another `Action` or `Predicate`. Interestingly enough, the number of `targets` is considerably greater than the number of `subjects` because the target role is often associated with a greater variety of grammatical roles. Furthermore, `is-a` is the most frequent relation in a corpus, appearing in 54.16% of sentences. This relation is relevant because it enables automatic building of ontological hierarchies. Another relevant relation, specifically in the medical domain, is `causes`, which appears in a 34.12% of sentences. This relation permits inference drawing to produce new knowledge from existing information.

To evaluate the quality of the corpus, we design an inter-annotator agreement metric. Even though Cohen's Kappa [43] is a common choice when evaluating inter-annotators agreement, in this case it is not convenient because this metric assumes a binary decision for each annotation whereas the *eHealth-KD v2* 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.

Hence, for the *eHealth-KD v2* corpus we define a metric that scores partial agreement proportional to the relative overlapping of the text spans. For each entity type E_t, relation type R_t, and and attribute type A_t, we define μE_t, μR_t, and μA_t respectively as the micro average of all the annotations corresponding to that specific entity, relation, and attribute between a pair of annotators, adding 1 for each pair of coincident annotations and a value $\delta(A, B)$, where $0 \leq \delta < 1$ for partial annotations between annotators A and B, equal to the relative number of coincident characters (equations 1, 2, and 3).

$$\mu(E_t) = \frac{\sum_{A,B \in E_t} \delta(A, B)}{|E_t|} \tag{1}$$

$$\mu(R_t) = \frac{Correct_t}{Correct_t + Invalid_t} \tag{2}$$

$$\mu(A_t) = \frac{Correct_t}{Correct_t + Invalid_t} \tag{3}$$

In the case of relations and attributes, the agreement score is computed among the subset of entities in which both annotators agree, otherwise the results would be unfairly skewed by disagreement among the entities. However, the overall micro-average does consider all of the annotations and thus provides a fair evaluation of the entire corpus.

Metric	Total	Trial	Training	Validation	Test
Sentences	1,045	45	600	100	300
Entities	6,612	292	3,818	604	1,898
`Concept`	4,092	181	2,381	368	1,162
`Action`	1,742	82	976	167	517
`Predicate`∗	563	27	330	45	161
`Reference`∗	215	2	131	24	58
Relations	6,049	232	3,504	537	1,776
`target`	1,729	88	974	166	501
`subject`	894	49	511	74	260
`in-context`∗	677	28	403	67	179
`is-a`	566	0	337	56	173
`in-place`∗	400	19	251	25	105
`causes`∗	367	0	219	27	121
`domain`∗	364	20	201	28	115
`argument`∗	343	16	201	28	98
`entails`∗	167	0	89	14	64
`in-time`∗	165	12	89	24	40
`has-property`	159	0	91	21	47
`same-as`	124	0	85	6	33
`part-of`	94	0	53	1	40
Attributes	585	28	311	69	177
`diminished`∗	18	1	8	2	7
`emphasized`∗	124	4	69	10	41
`negated`∗	164	4	94	24	42
`uncertain`∗	279	19	140	33	87

Table 3: Summary statistics for the *eHealth-KD v2* corpus. Labels marked with ∗ have been incorporated in this version of the corpus.

Table 4 summarizes the quality metrics. It considers each annotation type separately and combined, as well as the micro-average across all entity types and all relation types. Agreement scores are reported at different steps of the annotation process: between the non-expert annotators (stage 2), between non-expert annotators and the final revised version by the expert committee (stage 3) and between the merged version and the final version (stage 4). The most informative of these metrics is at stage 3, since it compares the annotation of non-expert humans with the final published annotations of the corpus. The stage 4 metric is provided only to illustrate that the expert revision of the normalized annotations produced minimal changes. Overall, the annotation agreement in eHealth-KD v2 is relatively high.

Agreement	**Stage 2**	**Stage 3**	**Stage 4**
Entities μE	0.7050	0.8159	0.9854
μE_{Action}	0.6989	0.8011	0.9892
$\mu E_{Concept}$	0.7810	0.8737	0.9929
$\mu E_{Predicate}$	0.4324	0.6641	0.9569
$\mu E_{Reference}$	0.7315	0.7990	0.9390
Relations μR	0.5146	0.7162	0.9692
μR_{arg}	0.6053	0.7782	0.9592
μR_{causes}	0.4006	0.6465	0.9917
μR_{domain}	0.6530	0.8004	0.9761
$\mu R_{entails}$	0.1030	0.4321	0.9623
$\mu R_{has-property}$	0.3684	0.6007	0.9737
$\mu R_{in-context}$	0.4195	0.6499	0.9584
$\mu R_{in-place}$	0.4165	0.6497	0.9407
$\mu R_{in-time}$	0.3677	0.6151	0.9346
μR_{is-a}	0.5439	0.7373	0.9750
$\mu R_{part-of}$	0.3016	0.5000	0.8710
$\mu R_{same-as}$	0.4662	0.6641	0.9242
$\mu R_{subject}$	0.5469	0.7294	0.9784
μR_{target}	0.6574	0.8139	0.9821
Attributes μA	0.4663	0.6537	0.9499
$\mu A_{diminished}$	1.0000	1.0000	1.0000
$\mu A_{emphasized}$	1.0000	1.0000	1.0000
$\mu A_{negated}$	0.9746	0.9888	1.0000
$\mu A_{uncertain}$	0.9370	0.9742	1.0000
Global agreement μ	0.6190	0.7667	0.9765

Table 4: Summary of the inter-annotator agreement score at different stages of the annotation process, for all entity and relation types.

4. The eHealth Knowledge Discovery Task

In this section we propose a formal definition for a knowledge discovery task based on the annotated corpus (Section 4.1), as well as evaluation metrics to allow an objective comparison

between different approaches (Section 4.2). Additionally, a computational infrastructure is provided that automates the evaluation process (Section 4.3).

4.1. Tasks and Evaluation Scenarios

Overall, the task consists of automatically identifying the annotated elements, i.e., entities and relations, in the *test* collection of the corpus. To evaluate a specific solution to this task, researchers are expected to use only the *training* set for learning model parameters and the *validation* set for adjusting hyper-parameters. Obviously, the *test* set must only be used for calculating the evaluation metrics. No model tuning or design decisions should be based on the output of these metrics, to avoid overfitting in the *test* set. Notice that we purposefully ignore the attributes in this task since entity and relation extraction is already a sufficiently complex challenge.

To better evaluate the strengths and weaknesses of different approaches, the annotation task is divided into two subtasks:

Subtask A: Entity recognition. The purpose of this subtask is to identify all the entities mentioned in a sentence and their corresponding classes (i.e., `Concept`, `Action`, `Predicate` and `Reference`.

Subtask B: Relation extraction. The purpose of this subtask is to detect all semantic relations between every pair of entities already labeled in each sentence.

This division into two subtasks does not necessarily mean that any given solution must explicitly solve both subtasks separately. Although this approach is the most commonly applied so far, there is evidence that end-to-end approaches solving both entity recognition and relation extraction can outperform approaches that solve both subtasks sequentially (see Section 4.2). However, since errors during entity recognition will necessarily translate into missing or spurious relations, splitting the evaluation into two subtasks allows for a more fine-grained evaluation of a given solution. Hence, in order to evaluate both end-to-end approaches and sequential approaches, we propose dividing the *test* set into three subsets of 100 sentences each, and perform three distinct evaluation scenarios respectively.

Scenario 1: End-to-end evaluation. In this scenario, both subtasks are evaluated. The input only consists of a plain text file with 100 sentences. Both end-to-end approaches and sequential approaches can be evaluated.

Scenario 2: Entity recognition evaluation. In this scenario only subtask A is evaluated. The input consists of plain texts but the expected output only requires entity annotations. This scenario also allows researchers to evaluate approaches that only perform entity recognition.

Scenario 3: Relation extraction evaluation. In this scenario, only subtask B is evaluated. The input consists of plain text and all the corresponding gold annotations for entities. The expected output consists of all the semantic relations occurring only between the annotated entities. This scenario allows researchers to evaluate approaches that only perform relation extraction and which require entities already annotated.

4.2. Evaluation Metrics

We propose an extended version of the F_1 metric modified to deal with partial matches to evaluate both subtasks. The F_1 metric depends on micro-averaging correct, incorrect, partial, missing and spurious annotations across the entire *test* set. Depending on the subtask(s) under evaluation, we define the following types of outcomes:

Subtask A - Correct C_A: when an annotation matches exactly with the corresponding gold annotation.

Subtask A - Incorrect I_A: when an annotation matches with a gold annotation with respect to the text span but defines a different entity label.

Subtask A - Partial P_A: when a text span has a non-empty but inexact intersection with a gold annotation, such as the case of "*respiractory tract*" and "*tract*" in Figure 1, Sentence 5. Partial phrases are only matched against a single correct phrase (i.e., the first partially matching phrase starting from the beginning of the sentence) to prevent a few large text spans that cover most of the document from getting a very high score.

Subtask A - Missing M_A: when an annotation that appears in the gold collection is not produced.

Subtask A - Spurious S_A: when an annotation is produced that does not appear in the gold collection.

Subtask B - Correct C_B: when a relation between two entities exists in the gold collection.

Subtask B - Missing M_B: when a relation in the gold collection is not produced.

Subtask B - Spurious S_B: when a relation is produced but it does not appear in the gold collection.

We define *Precision*, *Recall*, and F_1 as usual, taking into consideration that for each evaluation scenario only the terms related to the subtask(s) under evaluation are considered.

$$Precision = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + S_A + S_B} \tag{4}$$

$$Recall = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + M_A + M_B} \tag{5}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

Finally, we propose F_1 as defined in Equation 6 as the official metric to compare different approaches. If approaches are compared with respect to solving the complete task, then the first 100 sentences of the set (i.e., scenario 1) should be used. Otherwise, the second or third subset of 100 sentences should be used respectively to evaluate subtask A or B.

4.3. Evaluation Infrastructure and Baselines

To support researchers in the development of knowledge discovery technologies, we provide a toolkit and infrastructure that enables a faster and more objective experimentation process. These resources are freely available for the research community in a collection of Gitlab repositories[12].

The toolkit consists of the following elements:

- Plain text files and annotations in BRAT Standoff format for the *eHealth-KD v2* corpus, divided into the 4 collections described in Section 3.2.
- Configuration files necessary for deploying a BRAT server to explore and extend the *eHealth-KD v2* corpus, or to create other linguistic resources based on the annotation model described in Section 3.1.
- Utility scripts in the Python programming language for loading and manipulating the BRAT Standoff annotations in a computationally suitable format, as well as for producing correctly formatted output.
- Scripts for setting up and running an evaluation pipeline for the task defined in Section 4.2, including the three defined scenarios, and computing the official evaluation metrics.
- A set of baseline implementations with different degrees of complexity, including a random baseline and several classic machine learning approaches.

Using the aforementioned tools, researchers can quickly develop new approaches by extending the provided baselines, or developing a solution from scratch, without having to deal with setting up the evaluation environment or implementing the evaluation metrics. Furthermore, besides being able to evaluate their solutions offline, researchers can also upload their solutions to a cloud evaluation environment and automatically obtain the relevant metrics as well as compare their results with solutions already published. An official leaderboard is maintained that serves as an up-to-date state-of-the-art in all of the tasks. This information contains not only results, but also structured information about the approaches used and links to the relevant publications[13].

	Score (F_1)		
Team	**End-to-end**	**Subtask A**	**Subtask B**
Human	0.727	0.861	0.735
Dummy	0.424	0.546	0.123
Random	0.116	0.205	0.014

Table 5: Results (F_1 metric) of the baseline strategies in each scenario.

For reference purposes, Table 5 summarizes the results obtained by the different baselines implemented. The Dummy baseline learns all entities and relations that occur in the training set and builds a map relating tokens to entity labels, and token pairs to relation labels. This classifier uses the learned mappings to predict in the test set, hence only the words that appear in the

[12] https://ehealthkd.gitlab.io
[13] https://ehealthkd.gitlab.io/results

training set are recognized, which results in a significantly lower recall than precision. For the relations, the classifier uses the same strategy analyzing pairs of words. The Random baseline simply outputs a random label for each entity and relation pair, based on their relative frequency of appearance in the training set.

A human baseline is also provided for comparison purposes. To compute this baseline, one of the original participants in the annotation (a human expert) was invited to annotate the test collection, six months after the original annotation campaign. The results show that Subtask A is considerably easier both for humans and automatic systems than Subtask B. However, the difference between the algorithmic and human baselines is considerably larger in Subtask B, which indicates a larger margin for improvement.

5. The eHealth Knowledge Discovery Challenge

The *eHealth-KD v2* corpus was chosen for a shared competition presented at the IberLEF 2019 workshop, where the task was to design a computational system that can automatically provide the right annotations for a plain text input, as described in Section 4. The results of the competition are presented in detail in Piad-Morffis et al. [40]. Table 6 summarizes the results obtained by all participant systems in the competition.

To simplify the comparison and better understand the characteristics of each system, we define several tags to describe the kind of techniques used in each approach: (**C**)onditional (**r**)andom fields; (**P**)retrained or (**C**)ustom word (**e**)mbeddings; (**Ch**)aracter-level embeddings; hand-crafted (**R**)rules; natural language processing (**F**)eatures; dealing with the (**O**)verlapping of entities; (**At**)tention mechanisms; (**Co**)nvolutional layers; dataset (**Au**)gmentation techniques; and, if they solve both subtasks in a (**J**)oint form rather than separated.

Team		**Techniques**	**Score (F_1)** **End-to-end**	**Subtask A**	**Subtask B**
Human			0.727	0.861	0.735
Baseline (b)		**R**	0.430	0.546	0.123
TALP-UPC	[44]	**Cr-Pe-F-O-At-J-Au**	**0.639**	**0.820**	**0.626**
coin_flipper	[45]	**Pe-R-F**	**0.621**	0.787	**0.493**
LASTUS-TALN	[46]	**Cr-Ce-F-At**	**0.581**	**0.816**	0.229
NLP_UNED	[47]	**Pe-F-At**	0.547	0.754	**0.533**
HULAT-TaskAB	[48]	**Cr-Pe-Ch-Au**	0.541	0.775	0.123^{b}
UH-Maja-KD	[49]	**Cr-Ce-Ch-R-F-O**	0.518	**0.815**	0.433
LSI2_UNED	[50]	**Pe-Ch-F-Co**	0.493	0.731	0.123^{b}
IxaMed	[51]	**Cr-Ce-F-At**	0.486	0.682	0.435
HULAT-TaskA	[52]	**Cr-Pe-Ch-Au**	0.430^{b}	0.790	0.123^{b}
VSP	[53]	**-**	0.428^{b}	0.546^{b}	**0.493**

Table 6: Results (F_1 metric) in each scenario, sorted by Scenario 1 (column *Score*). The top results per scenario are highlighted in **bold**. Results that use the baseline implementation are represented by $\#^{b}$. The dummy baseline implementation provided in the challenge is slightly different due to variations in the order of the training sentences with respect to Table 5. Adapted from Piad-Morffis et al. [40].

In terms of modeling, most approaches tackle both subtasks sequentially, feeding the output of subtask A to the pipeline for solving subtask B. The most natural representation of subtask A

presented in the challenge is as a sequence labeling problem. Several approaches deal with the problem of overlapping entities using BIOUV tags or similar encoding systems. Afterwards, the most common computational solution for the labeling problem consisted of some variation of LSTM or Bi-LSTM architectures, commonly followed by a final CRF layer. In terms of features, some approaches introduced domain-specific word embeddings trained in selected corpora, but most resort to Glove or Word2Vec. Several approaches include also character-level embeddings to deal with morphological phenomena and part-of-speech tags to capture the grammatical structure of the sentence.

For subtask B, the most common modeling consists of considering then N^2 entity pairs in each sentence as separate classification problems. Some approaches build a single model with multi-label output while others build separate binary models for each label. In terms of features, the same token-level linguistic and morphological features are used, along with one-hot encoded or embedded entity classes. One approach (*UH-Maja-KD* [49]) trains a recurrent network for encoding the path between each pair of entities in the dependency tree, using POS-tags as features. In contrast with most of the approaches, the best performing system in all three scenarios (*TALP-UPC* [44]) presents a unified architecture that solves both subtasks simultaneously.

5.1. Analysis of Systems Approaches

To evaluate the impact of each design component (e.g., using CRF, embeddings, etc.) on the overall performance of each system, a linear regression model is fitted on the challenge results. Each system is represented as the set of tags corresponding to the techniques used in that system, as described in Table 6, second column. The linear regression model assigns a weight to each of the tags that approximates its relative impact when considering all the systems in which that tag is present, see Table 7. For example, tag **At**, which corresponds to the use of attention-based architectures, obtains a score of 0.141 for Subtask B. This indicates that, all other things considered equal, if a system utilizes this type of technique we can expect an average increase in F1 score of 0.141 in Subtask B, compared with not using this technique but maintaining all the remaining characteristics. The weights computed for each technique are only an approximation of its relative importance, since this analysis assumes independence between the techniques used, which is obviously not a realistic assumption. However, the R^2 score for the main scenario is 0.773, and for the other two subtasks is 0.857 and 0.936 respectively, which indicate that these weights provide an adequate estimation of the impact of each technique, especially for Subtasks A and B.

As expected, one of the most significant factors for increasing performance in the end-to-end scenario (Scenario 1) is solving both tasks simultaneously. The only system that applies this strategy obtains the best results and the linear regression weights are relatively higher. Using NLP features in addition to word embeddings and performing dataset augmentation also provide a significant boost to performance, possibly given the relatively small size of the training set in comparison with the task complexity. An additional positive effect is caused by the use of custom rules, such as *coin_flipper*'s strategy for merging entities [45]. Counter-intuitively, the use of custom word embeddings produces a marginally negative effect, presumably given the difficulty of learning embeddings on domain-specific text, where it is difficult to obtain a sufficiently large corpus.

Specifically for subtask A (Scenario 2), the strategies that provide marginal advantages are related to handling overlapping and discontinuous entities. This is an indication that most systems are able to correctly deal with the "easier" instances, i.e., single-word entities or continuous

Technique		Scenario		
		End-to-end	Subtask A	Subtask B
Attention-based architecture	**(At)**	-0.015	-0.002	**0.141**
Character embeddings	**(Ch)**	-0.088	-0.006	-0.129
Convolutional networks	**(Co)**	0.019	-0.018	-0.140
Conditional random fields	**(Cr)**	0.010	0.011	-0.103
Custom embeddings	**(Ce)**	-0.012	-0.008	-0.087
Dataset augmentation	**(Au)**	**0.022**	**0.019**	-0.016
Hand-crafted rules	**(R)**	**0.059**	**0.031**	**0.101**
Joint solution (end-to-end)	**(J)**	**0.042**	0.015	0.081
NLP features	**(F)**	0.021	-0.004	0.021
Overlapping entities	**(O)**	-0.002	**0.039**	**0.270**
Pretrained embeddings	**(Pe)**	0.012	0.008	0.010

Table 7: Relative impact of the characteristics of each system in the overall score, per scenario, as defined by a linear regression model fitted on each system's performance. Tag labels correspond to the techniques used by each system as reported in Table 6. Highlighted in **bold** are the most significant weights in each scenario. Adapted from Piad-Morffis et al. [40].

entities with no overlap, and thus it is in the remaining cases where differences occur. In subtask B (Scenario 3) the overlapping subproblem is also relevant, presumably because otherwise a large number of missing relations would be reported. The use of attention mechanisms also provides a positive boost, in contrast with previous scenarios,presumably because it helps to capture long-range dependencies between entities that are far apart in a sentence.

By far the most significant factor that influences the correct identification of each entity and relation type is the number of instances in the training set. To illustrate this insight, Figure 4 plots the relative number of instances of each annotation identified by at least one system in relation to their frequency in the training set. The most significant deviation from the $y = x$ line is the `Reference` entity type, which by design is mostly characterized by a relatively short number of linguistic constructions, easily recognizable by part-of-speech tags.

The previous analysis can shed light on the type of techniques that are more promising for solving the eHealth-KD task. However, these results must be weighted with caution since the analysis is based on simplistic assumptions, such as independence of each technique. A simple linear model is unlikely to capture the complex interactions between components in a knowledge-discovery system. Nevertheless, some high-level insights can be extracted from this analysis. First, the use of specific techniques seems to have a larger impact on Subtask B, where weights have a higher variance, than on Subtask A. This could indicate that Subtask A is generally solvable with a larger variety of techniques, while Subtask B requires a more careful design. And second, even though modern deep learning techniques are the go-to approach in NLP, complex tasks like relation extraction still require taking into consideration phenomena like the overlapping of entities, which involve hand-crafted rules. Applying black-box deep learning architectures without considering these type of intricacies is unlikely to yield state-of-the-art performance.

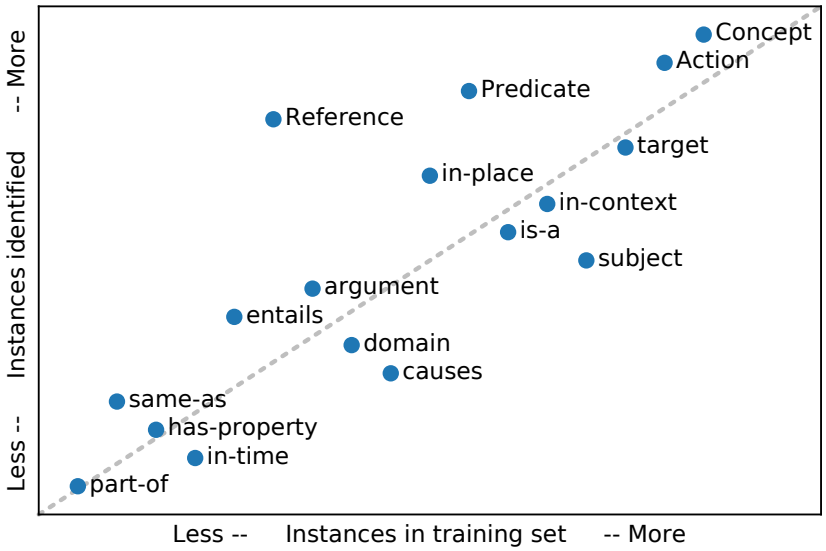

Figure 4: Correlation between the number of instances identified by one or more systems, and the relative frequency of labels in the training set. Adapted from Piad-Morffis et al. [40].

6. Overall Discussion

This section presents an overall discussion of the main takeaways of this research, the lessons learned and the limitations of the current solutions proposed for the eHealth Knowledge Discovery Task. We also highlight interesting ideas for further exploration and research, based on insights obtained by analyzing the most promising approaches.

Fostering Research in eHealth Knowledge Discovery

Different semantic representations for capturing knowledge expressed in natural language have been developed (e.g. AMR, FrameNet and PropBank). The main drawback of these representations is their complexity, since they often depend on lexicons that define the specific semantic roles for each word. Thus, developing artificial intelligence systems for knowledge discovery with this level of detail is a very challenging problem. Using simpler semantic representations that do not rely on word-specific roles or relations, even at the cost of reducing expressibility, can simplify the development of automatic techniques based on machine learning.

This research proposes a line of development in this direction, whereby knowledge discovery with a high level of abstraction can be subsequently refined for domain specific tasks. The purpose is not to replace fine-grained semantic representations, such as AMR or FrameNet, but rather to provide a more coarse-grained representation that can be used as an initial step in different knowledge discovery tasks. This type of semantic representation can potentially aid in downstream tasks like ontology learning, in the same way that general-purpose POS-tagging is often performed prior to more complex NLP tasks like question answering.

The resources, tools and infrastructure developed in this research aim to provide a foundation for the research community to build such general-purpose semantic representation techniques. Succeeding in this endeavor will depend not only on theoretical advancements such as better deep learning architectures or natural language processing techniques, but also on the availability

of resources that enable an efficient experimentation. In this sense, our proposal introduces a new knowledge discovery task together with formally defined evaluation metrics, as well as a practical test-bed where researchers can quickly develop new techniques and obtain immediate feedback. It is also a step in the direction of encouraging knowledge discovery research in less developed languages, such as Spanish, and in socially important domains such as health.

Current and Future Challenges

The results of the eHealth-KD Challenge shed some light on the complexity of the various steps involved in the design of automatic knowledge discovery systems for this task. Most of the systems modeled the task as a pipeline in which first entities are recognized and then relations are extracted. The entity recognition part was commonly modeled as a sequence tagging problem and solved by standard techniques, e.g., Bi-LSTM networks and Conditional Random Fields. The relation extraction part was commonly modeled as a standard classification problem, where the input consists of some sensible representation of a pair of concepts, using context-aware embeddings and other syntactic features. In contrast, the best performing system of the challenge consists of an end-to-end approach that outputs both entity types and relevant relations for each pair of potential concepts detected in each sentence. Besides marginal differences in architecture and training methodology, we argue that this system's strength arises from the regularizing effect of learning a unified representation for both subtasks, instead of different representations, which allows it to obtain more information from the same amount of training data. Furthermore, using the previous version of the corpus provided it with some leverage in terms of increasing statistical coverage.

Based on these observations, we estimate that successful approaches to this problem should consider the following strategies: solving both problems simultaneously rather than sequentially; using general-purpose pre-trained word embeddings rather than customized ones; applying some form of dataset augmentation to increase statistical coverage; and, designing problem-specific rules to deal with overlapping and discontinuous entities.

In comparison with the human baseline, subtask B appears to be considerably harder for machine learning systems than for humans. Specifically, the human baseline beats the best performing system by an absolute 8.8% in the full scenario, but only by a 4.1% in subtask A, compared with a 10.9% in subtask B. Intuitively, subtask B should be harder, since the number of labels to predict is larger than in subtask A. However, this does not explain the difference in performance between humans and machine learning systems. On average, the systems that attempt to solve subtask B obtain an often significantly lower F_1 in subtask B compared to the full scenario (Scenario 1), while the human baseline is slightly better at subtask B. This indicates that humans can leverage some additional insights by seeing the correct annotations for subtask A that machine learning systems fail to recognise. However, the fact that subtask B is significantly harder for humans than subtask A is an indication of the high degree of qualitative analysis involved in this problem. As such, there is a threshold above which even human experts will not completely agree, given the inherently subjective nature of natural language understanding.

In the light of these considerations, we believe there is still a large margin for improvement via a more principled approach that considers the global information of the complete sentence rather than simplifying the problem as a set of unconnected classification subtasks, one for each pair of tokens. From a human perspective, the annotation of a sentence is a global process, in which a decision to consider a specific word as an `Action` or `Predicate` makes an annotator reconsider the whole sentence and potentially change other annotations. Incorporating this type of global awareness into a system requires more than just applying context-aware embeddings

or even sentence-level language models. The system must be able to assess an incompletely annotated sentence and potentially undo or correct previous labels as it progresses, until a suitable convergence criteria is reached. This kind of behavior requires a more expressive framework than that offered by pure supervised learning architectures. A possible approach involves some sort of annotator agent that observes the complete sentence, and performs actions akin to how humans approach this problem. We believe that dependency parsers are a good starting point in devising such an architecture.

Another important consideration is the degree to which the successful identification of each entity and relation label is correlated with its frequency in the training set (see Figure 4). This reinforces the idea that most current approaches are basically performing pure statistical learning and thus, are not capable of accurately capturing the semantic nuances of each of these labels. This evidence also points to the necessity of more principled approaches that actually attempt to understand the semantic meaning of the annotation model rather than simply learning by statistical association. Given that producing human annotated resources with this level of fine-grained semantics is time consuming, it is unlikely that pure statistical approaches will ever be sufficient to learn in this context by supervised training alone.

Notwithstanding this, the emergence of Transformer architectures and their recent success at several NLP tasks [54] opens the door to potentially improving current results in the eHealth-KD challenge with little additional effort. The first edition of the challenge (in 2018) consisted mostly of hybrid systems, using a combination of rule-based and knowledge-based NLP techniques with machine learning. However, the 2019 edition of the challenge included almost no rule-based approach, in favor of more complex deep learning architectures. It is likely that future editions of the challenge will see the rise of Transformers as the leading technology, potentially combined with problem-specific architecture designs, such as the ones discussed previously.

Existing Limitations

Compared to similar work and a previous version of this corpus, our main focus in this research has been related to increasing the expressibility of composite concepts. The previous version of the corpus allowed for composite concepts via the annotation of `Actions` and their corresponding roles. This research introduces `Predicates` and `Contextual relations` that allow for a finer semantic representation when composing complex concepts. Additionally, we introduce `causality` and `entailment` as two specific relations with well-defined semantics. These types of relations could enable the construction of inference systems which can discover new knowledge by the successive application of inference rules, given that `causality` and `entailment` are transitive relations

Our work has so far focused on Spanish language, given the relatively lower predominance of Spanish-bashed resources compared to English ones. However, the annotation model has been designed with the explicit objective of being applicable across many languages. The core elements are all language-agnostic. This is because concepts, actions, references and predicates, as well as the semantic relations defined, are found in all human languages, even if their syntactic representation is different. Our model explicitly avoids syntactic rules-of-thumb in favor of purely semantic definitions. For example, a common mistake found in early annotators was unconsciously tagging verbs as actions. Even though this might be correct in many sentences, we explicitly forbade such rules to avoid biasing annotators towards a syntactically-based mind-set. Likewise, the definition of attributes (`uncertain`, `diminished`, `emphasized` and `negated`) is an effort to generalize several distinct grammatical patterns into semantic annotations independent of the surface text. However, it is still an open question as to whether our annotation

model will generalize successfully to other languages. An early proof-of-concept is being actively developed at the time of writing, with the successful annotation of English research papers on the subject of the COVID-19 pandemic, using the same annotation model proposed in this research.[14]

Increasing the expressibility of an annotation model also introduces new sources for ambiguity. During the annotation process, we discovered this to be a major source of inter-annotator disagreement, especially when deciding between `Predicate` and `in-context`. Another source of ambiguity was detected in the different semantic roles assigned to `target` annotations. One of such roles is similar to ConceptNet's `MotivatedByGoal` and `UsedFor`, i.e., to indicate that an `Action` is performed with a purpose. This usage is different to `causes` and `entails` and might require the addition of a new semantic relation. As a final remark, although entity attributes are accounted for in the annotation model and included in the corpus, they are excluded from the evaluation since they add significant complexity to an already challenging computational task. However, in future editions of the eHealth-KD challenge they will be evaluated, possibly as part of an additional scenario.

7. Conclusions and Future Work

This research presents the design and construction of an ecosystem for the development of knowledge discovery technologies in the biomedical domain. This ecosystem includes linguistic resources, computational tools and a methodology for the evaluation of new approaches. An annotation model was defined to capture the most relevant semantic content of natural language sentences, based on Subject-Action-Target tuples and additional semantic relations. The model does not include domain-specific entities or relations so as to be as general as possible. Based on this model, a corpus of 1,045 sentences in the Spanish language was manually annotated, taken from an online source of health information. The corpus enables the construction of fine-grained knowledge discovery systems that can be applied in multiple domains. With this purpose in mind, a shared evaluation campaign was organized, in which 10 teams of researchers proposed different strategies, mostly focused on deep learning architectures that achieved significant results.

To foster continued development in this line of research, an infrastructure and toolkit for researchers is made available, including baseline implementations, an ongoing evaluation environment in the cloud, and up-to-date statistics on the state-of-the-art in the aforementioned task. These results build on previous research and are part of a continued attempt to leverage general-purpose semantics and knowledge-based technologies together with novel deep learning architectures for the construction of automatic knowledge discovery technologies.

Acknowledgments

Funding: This research has been partially supported by the University of Alicante and University of Havana, the Generalitat Valenciana (*Conselleria d'Educació, Investigació, Cultura i Esport*) and the Spanish Government through the projects SIIA (PROMETEO/2018/089, PROMETEU/2018/089) and LIVING-LANG(RTI2018-094653-B-C22).

[14] https://github.com/matcom/cord19-ann

References

[1] Richard M Goldberg, John Mabee, Linda Chan, and Sandra Wong. Drug-drug and drug-disease interactions in the ed: analysis of a high-risk population. *The American journal of emergency medicine*, 14(5):447–450, 1996.

[2] Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):S3, 2005.

[3] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[4] Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.)*, 2018.

[5] S Estevez-Velarde, Y Gutierrez, A Montoyo, A Piad-Morffis, R Munoz, and Y Almeida-Cruz. 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, Computer ..., 2018.

[6] David Crystal. *The Cambridge encyclopedia of the English language*. Ernst Klett Sprachen, 2004.

[7] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, 2013.

[8] Alejandro Piad-Morffis, Yoan Gutiérrez, Yudivian Almeida-Cruz, and Rafael Muñoz. [dataset] eHealth-KD v2, March 2020. URL https://doi.org/10.5281/zenodo.3696792.

[9] Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza, and Arantza Casillas. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56:318 – 332, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2015.06.016. URL http://www.sciencedirect.com/science/article/pii/S1532046415001264.

[10] Isabel Moreno, Ester Boldrini, Paloma Moreda, and M Teresa Romá-Ferri. Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of biomedical informatics*, 72:8–22, 2017.

[11] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5): 914–920, 2013.

[12] BioAMR Corpus, 2018. URL https://amr.isi.edu/download/2018-01-25/amr-release-bio-v3.0.txt.

[13] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.

[14] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[15] Alejandro Piad-Morffis, Yoan Gutiérrez, and Rafael Muñoz. A corpus to support ehealth knowledge discovery technologies. *Journal of Biomedical Informatics*, 94:103172, 2019. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2019.103172. URL http://www.sciencedirect.com/science/article/pii/S1532046419300905.

[16] Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. Propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*, 2010.

[17] Noelia Migueles-Abraira, Rodrigo Agerri, and Arantza Diaz de Ilarraza. Annotating abstract meaning representations for spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[18] Mariana Neves and Jurica Ŝeva. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 12 2019. ISSN 1477-4054. doi: 10.1093/bib/bbz130. URL https://doi.org/10.1093/bib/bbz130. bbz130.

[19] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi 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*, pages 102–107. Association for Computational Linguistics, 2012.

[20] Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. A web-based tool for the integrated annotation of semantic and syntactic structures.

In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, 2016.

[21] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029, 2013.

[22] Philip Ogren. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 273–275, 2006.

[23] Dongseop Kwon, Sun Kim, Soo-Yong Shin, Andrew Chatr-aryamontri, and W John Wilbur. Assisting manual literature curation for protein–protein interactions using bioqrator. *Database*, 2014, 2014.

[24] Jan Christoph Meister, Jan Horstmann, Marco Petris, Janina Jacke, Christian Bruck, Mareike Schumacher, and Marie Flüh. Catma, October 2019. URL https://doi.org/10.5281/zenodo.3523228.

[25] ExplosionAI GmbH. Prodigy, 2017-2020. URL https://prodi.gy/.

[26] Jin-Dong Kim, Yue Wang, Shigeru Nakajima, and Nakashima Masahiro. TextAE, 2018. URL http://github.com/pubannotation/textae.

[27] LightTAG. LightTAG the text annotation tool for teams, 2018. URL https://www.lighttag.io.

[28] Emilia Apostolova, Sean Neilan, Gary An, Noriko Tomuro, and Steven Lytinen. Djangology: A light-weight web-based tool for distributed collaborative text annotation. 2010.

[29] David Salgado, Martin Krallinger, Marc Depaule, Elodie Drula, Ashish V Tendulkar, Florian Leitner, Alfonso Valencia, and Christophe Marcelle. Myminer: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, 28(17):2285–2287, 2012.

[30] Johannes Kiesel, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Wat-sl: a customizable web annotation tool for segment labeling. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 13–16, 2017.

[31] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the share/clef ehealth evaluation lab 2013. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40802-1.

[32] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martinez, Guido Zuccon, and João Palotti. Overview of the share/clef ehealth evaluation lab 2014. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 172–191, Cham, 2014. Springer International Publishing. ISBN 978-3-319-11382-1.

[33] Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. Overview of the pan/clef 2015 evaluation lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 518–538. Springer, 2015.

[34] Aurélie Névéol, K Bretonnel Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier, et al. Clinical information extraction at the clef ehealth evaluation lab 2016. In *CEUR workshop proceedings*, volume 1609, page 28. NIH Public Access, 2016.

[35] Jonathan May and Jay Priyadarshi. Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, 2017.

[36] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*, 2017.

[37] Rafael Glauber. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. 2019.

[38] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodrguez, JA Lopez Martin, Marta Villegas, and Martin Krallinger. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS. org), Bilbao, Spain (Sep 2019), TBA*, 2019.

[39] Eugenio Martínez Cámara, Yudivian Almeida Cruz, Manuel Carlos Díaz Galiano, Suilan Estévez-Velarde, Miguel Ángel García Cumbreras, Manuel García Vega, Yoan Gutiérrez, Arturo Montejo Ráez, Andres Montoyo, Rafael Muñoz, et al. Overview of tass 2018: Opinions, health and emotions. 2018.

[40] Alejandro Piad-Morffis, Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, and Andrés Montoyo. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS. org*, 2019.

[41] Alejandro Piad-Morffis, Yoan Guitérrez, Suilan Estevez-Velarde, and Rafael Muñoz. A General-Purpose Annota-

tion Model for Knowledge Discovery: Case Study in Spanish Clinical Text. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 79–88, 2019.

[42] Fausto Giunchiglia and Mattia Fumagalli. Teleologies: Objects, actions and functions. In *International conference on conceptual modeling*, pages 520–534. Springer, 2017.

[43] Susana M Vieira, Uzay Kaymak, and João MC Sousa. Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems*, pages 1–8. IEEE, 2010.

[44] Salvador Medina and Jordi Turmo. Talp-upc at ehealth-kd challenge 2019: A joint model with contextual embeddings for clinical information extraction. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[45] Neus Català and Mario Martin. coin_flipper at ehealth-kd challenge 2019: Voting lstms for key phrases and semantic relation identification applied to spanish ehealth texts. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[46] Alex Bravo, Pablo Accuosto, and Horacio Saggion. Lastus-taln at iberlef 2019 ehealth-kd challenge: Deep learning approaches to information extraction in biomedical texts. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[47] Hermenegildo Fabregat, Andres Duque, Juan Martinez-Romo, and Lourdes Araujo. Nlp_uned at ehealth-kd challenge 2019: Deep learning for named entity recognition and attentive relation extraction. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[48] Cristóbal Colón-Ruiz and Isabel Segura-Bedmar. Hulat-taskab at ehealth-kd challenge 2019: Knowledge recognition from health documents by bilstm-crf. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[49] Jorge Mederos-Alvarado, Ernesto Quevedo-Caballero, Alejandro Rodríguez-Pérez, and Rocío Cruz-Linares. Uhmaja-kd at ehealth-kd challenge 2019: Deep learning models for knowledge discovery in spanish ehealth documents. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[50] Alicia Lara-Clares and Ana Garcia-Serrano. Lsi2_uned at ehealth-kd challenge 2019: A few-shot learning model for knowledge discovery from ehealth documents. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[51] Iakes Goenaga, Sergio Santana, Sara Santiso, Koldo Gojenola, Alicia Pérez, and Arantza Casillas. Ixamed at ehealth-kd challenge 2019: Using different paradigms to solve clinical relation extraction. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[52] Alejandro Ruiz-de laCuadra, Jose Luis Lopez-Cuadrado, Israel Gonzalez-Carrasco, and Belen Ruiz-Mezcua. Hulat-taska at ehealth-kd challenge 2019: Sequence key phrases recognition in the spanish clinical narrative. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[53] Víctor Suárez-Paniagua. Vsp at ehealth-kd challenge 2019: Recurrent neural networks for relation classification in spanish ehealth documents. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[54] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.