A General-Purpose Annotation Model for Knowledge Discovery: Case Study in Spanish Clinical Text

Alejandro Piad-Morffis[1], **Yoan Gutiérrez**[2], **Suilan Estevez-Velarde**[1], **Rafael Muñoz**[2]

[1]School of Math and Computer Science, University of Havana
{apiad,sestevez}@matcom.uh.cu
[2]Deparment of Software and Computing Systems, University of Alicante
{ygutierrez,rafael}@dlsi.ua.es

Abstract

Knowledge discovery from text in natural language is a task usually aided by the manual construction of annotated corpora. Specifically in the clinical domain, several annotation models are used depending on the characteristics of the task to solve (e.g., named entity recognition, relation extraction, etc.). However, few general-purpose annotation models exist, that can support a broad range of knowledge extraction tasks. This paper presents an annotation model designed to capture a large portion of the semantics of natural language text. The structure of the annotation model is presented, with examples of annotated sentences and a brief description of each semantic role and relation defined. This research focuses on an application to clinical texts in the Spanish language. Nevertheless, the presented annotation model is extensible to other domains and languages. An example of annotated sentences, guidelines, and suitable configuration files for an annotation tool are also provided for the research community.

1 Introduction

Knowledge discovery is a field of computer science that shows an accelerated growth in the past three decades. Advances in this area have been applied in many domains, from databases (Fayyad et al., 1996; Stahl et al.) to images (Lu et al., 2016) and natural language text (Carlson et al., 2010). Specifically in natural language text, this field is highly relevant in the biomedical and health domains, where it is used for performing tasks such as Named Entity Recognition (NER), Relationship Extraction and Hypothesis Generation, among others. (Simpson and Demner-Fushman, 2012). These tasks generally use annotated corpora for learning the characteristics that appear in the text and mapping them to knowledge structures. For each task, specific annotation models have been designed that focus on specific elements of the text. For example, in NER tasks is more important to focus on nominal phrases than other grammatical constructions.

Despite that these domain-specific tasks are different, most of them share common characteristics. For example, most tasks deal with the detection of relevant entities and their relations. Hence, promoting general-purpose annotation models would allow the design of reusable and cross-domain knowledge discovery techniques. In this line, several domain-independent semantic representations have been developed (e.g., AMR (Banarescu et al., 2013), PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998)). However, these representations rely heavily on fine-grained lexicons that define specific semantic roles for each word meaning. Therefore, developing knowledge discovery systems with this level of detail supposes great challenges. Using more coarse-grained semantic representation, even with the loss of some representational capacity, would simplify the creation of automatic techniques based on machine learning. This representation could also be used as the first stage in a pipeline for a domain-specific task, thus reusing resources and techniques in domains with few available resources.

This paper presents a general-purpose annotation model specifically designed to enable knowledge discovery techniques in biomedical text. This model represents the most relevant aspects of the semantic meaning of sentences in natural language, that allows the representation of the basic knowledge contained in a sentence. Even though this model is language-agnostic, we focus on Spanish text because is a less pervasive language than English in terms of computational resources available. However, this model can be applied to several western languages (e.g., English, French, Spanish, Portuguese) without change, be-

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 79–88
Minneapolis, Minnesota, June 7, 2019.

cause it doesn't rely heavily on the grammatical structure of the sentence. At the moment of writing, this model is being used to annotate a Spanish corpus of clinical text for a shared evaluation task[1]. Relevant configuration files and example annotated sentences are also published online[2].

The remainder of the paper is organized as follows: Section 2 presents a brief review of annotation models and related corpora in the health domain. Section 3 describes our proposal for a general-purpose annotation model with examples and highlights its key design decisions. Section 4 proposes a methodology for the annotation, normalization, agreement and evaluation of a corpus based on this annotation model. Finally, Section 5 provides preliminary conclusions and prospects of our proposal.

2 Annotation models for knowledge discovery

In this section we present a review of relevant annotation models from which we draw inspiration. We focus general-purpose annotation models 2.1 as well as on annotation models that have been applied to the health domain 2.2.

2.1 General-purpose annotation models

Several general-purpose semantic annotation models have been developed, that attempt to represent the semantics of a sentence beyond the syntactic structure. These models are loosely based on the Subject-Verb-Object grammatical structure that is pervasive in human language.

PropBank (Palmer et al., 2005) proposes a general purpose annotation schema, based on annotating predicates (verbs) as the main semantic constituents of a sentence. ProbBank's annotation schema is able to represent several semantic relations, including the agent that causes an action, the receiver of the effects of an action, time and location modifiers, and causal relationships. One key characteristic of PropBank is that every predicate defines custom semantic roles, i.e., the predicate "*accept*" defines roles for the agent who accepts (`ARG0`), the object that is accepted (`ARG1`), and the agent from whom that object is accepted.

FrameNet (Baker et al., 1998) is a lexical database and an annotated corpus that models the semantic roles and relations in a natural language sentence through conceptual structures named *frames*. Frames represent general-purpose concepts, or events, that define the possible semantic relations in which those concepts can be realized in natural language.

VerbNet (Schuler, 2005) is a verb lexicon that also defines specific semantic roles for each verb. In VerbNet, verbs are organized in a hierarchy, and linked through different thematic roles, such as agents, cause, source, or topic. These elements allow to capture the semantic representation of sentences.

PropBank semantic roles are similar to the thematic roles defined in VerbNet and frame elements in FrameNet. As such, there are resources that link these semantic structures (Palmer, 2009).

A more recent proposal is Abstract Meaning Representation (Banarescu et al., 2013, ARM). AMR constitutes a semantic representation schema for English sentences that also attempts to cover a wide range of semantic relations with a general-purpose model. AMR includes PropBank semantic roles, as well as coreference resolution within the same sentence, named entities and types, negation, and other modifiers in a graph structure that represents the meaning of a natural language sentence. However, even though AMR captures the full semantic meaning of a sentence, for the purpose of knowledge discovery it is still considerably abstract, and additional processing is necessary to extract concrete structures of knowledge (Rao et al., 2017).

The annotation model proposed in this research shares similarities from general-purpose semantic annotation models such as AMR and PropBank. In contrast to these resources, our model makes no distinction between different types of actions, which are loosely related to verbs, as explained in Section 3. Instead, we define two general-purpose roles, the agent that performs and action, and the receiver of the effects of the action. These roles roughly correspond to `ARG0` and `ARG1` respectively in PropBank, although in specific cases their semantic meaning might differ. This simplification is directed towards enabling the automation of the annotation process with the use of machine learning techniques. Another key difference of our model is the inclusion of general-purpose taxonomic relations (e.g, *hypernomy/hyponomy* and *meronym/holonym*) that are inferred from the sen-

[1] https://knowledge-learning.github.io/ehealthkd-2019/

[2] https://github.com/knowledge-learning/satr-ann

tence. These relations are directed towards easing the automatic construction of knowledge bases.

2.2 Annotations models in the health domain

Knowledge discovery tasks in the health domain are often supported by the construction of manually-annotated corpora. Several task-specific annotation models have been developed for this purpose. One example is the DrugSemantics corpus (Moreno et al., 2017) where product characteristics are annotated, and BARR2 (Intxaurrondo et al., 2018) which is concerned with biomedical abbreviations. Many corpora include specific types of named entities relevant to the medical domain, such as DDI (Herrero-Zazo et al., 2013) which annotates drugs and other substances. Other examples include i2b2 (Uzuner et al., 2010) which annotates medications, dosages and other details of drug administration and CLEF (Roberts et al., 2009) which annotate different types of conditions, devices and their results in specific clinical cases. Given the specificity of the annotated concepts, most of these resources are built by biomedical experts.

The previous examples are corpora helpful in designing techniques oriented towards narrow tasks, where the annotation model is specifically designed to only consider portions of the text relevant to the concepts of interests (i.e., medical entities, genes, etc.). An alternative approach that attempts to model a wide range of the semantics of a document is Bio-AMR (May and Priyadarshi, 2017). This corpus contains health-related sentences annotated with their AMR structure, a general-purpose semantic representation of natural text. Another relevant resource is BioFrameNet (Dolbey et al., 2006), an extension to FrameNet with specific semantic roles for the biomedical domain. A positive consequence of using general-purpose semantic annotations is that it doesn't necessarily require experts in biomedical areas to participate in the annotation process.

The eHealth-KD corpus (Martínez Cámara et al., 2018) attempts to achieve a middle ground by representing a broad range of knowledge with a simple annotation model based on Subject-Action-Target triplets and 4 additional semantic relations. However, after the annotation process several shortcomings were identified. One example is the necessity for including causality and entailment as explicit relations, rather than representing them through actions, given the importance of this type of assertions in medical texts. Likewise, the annotation lacks the ability to represent coreferences ("*this*", "*that*"), and for this reason many sentences cannot be fully annotated. Also, complex linguistic constructions that represent composite concepts (e.g., "*the patients that received treatment*") are difficult to annotate, especially when they participate in other relations. This paper extends the annotation model used by the eHealth-KD corpus with semantic elements used in general-purpose annotation models, such as AMR and PropBank. This extension allows solving the aforementioned issues and increases its representational power without adding an overly complex set of new semantic roles and relations.

3 Annotation model

In this section, we define an annotation model that attempts to represent the most relevant semantic relations in a natural language sentence. This model should avoid ambiguities as much as possible, such that different human annotators can agree with a high probability. The model needs to be expressive enough to capture relevant domain concepts and their interactions. It must also be able to represent complex concepts that are built by the combination of simpler concepts. This model is designed to aid in the construction of knowledge discovery systems. For this reason, it is necessary to detach the model representation as much as possible from the grammatical structure of sentences, and instead attempt to represent its semantic meaning.

With these objectives in mind, the annotation model proposed in this research is based on the Subject-Verb-Object grammatical structure present in western languages. However, since we are interested in annotating fragments of knowledge, the semantic role of annotated entities does not necessarily match the grammatical role. The main semantic roles of this model are `Concept` and `Action`, which are used to represent factual information about what is being done, by who, to whom. These structures can be contextualized with time, location, and other general circumstances. An additional semantic role named `Predicate` is used to build more complex conceptualizations from simpler ones. Finally, 6 specific semantic relations are used to represent general-purpose knowledge. The relations `is-a`,

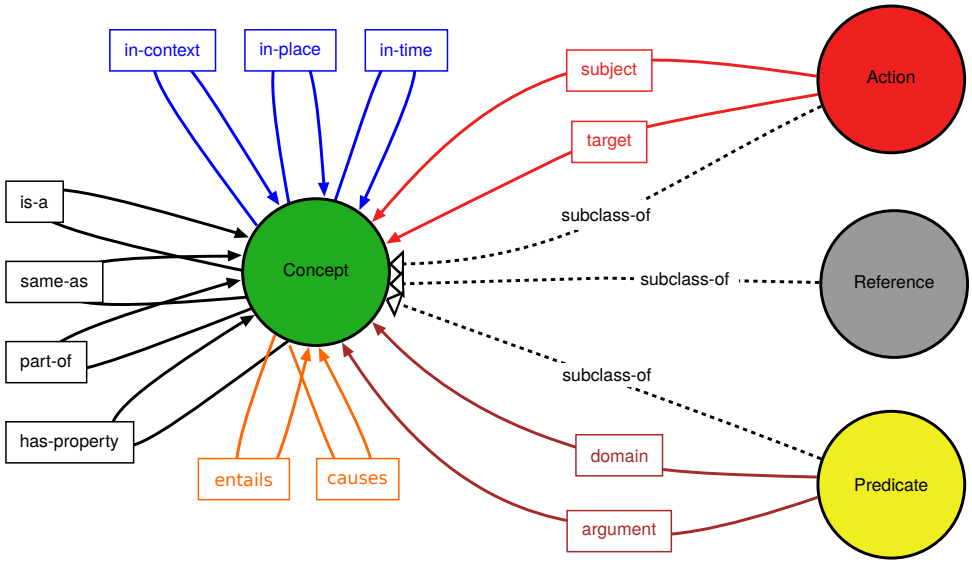

Figure 1: Conceptual schema for the annotation model. Each of the semantic roles defined in the annotation model are represented as circles. The possible relations defined between each pair of roles are represented in rectangles.

`same-as`, `has-property` and `part-of` are taken from taxonomic and ontologic representations, while the relations `causes` and `entails` are taken from the domain of text comprehension.

In contrast with AMR and PropBank, our annotation model does not yet specifies the semantic meaning of each `Concept` and `Action`. The actual meaning must be inferred from the text of the annotated entities. Likewise, the exact meaning of each semantic role (e.g, the receiver in "*accept*") is also inferred from the text, and must be resolver in a later stage. The taxonomic relations allow the capture of domain-specific definitions, that more in fine-grained task would be represented with specific entity types and relations. The domain-specific knowledge is thus represented by the semantic meaning of the annotated words, and not explicitly represented by specific entity types or relations.

The following sections explain each semantic role and relation in details and provide examples of its use in natural text sentences. Figure 1 shows a graphic representation of our annotation model.

3.1 Concepts

A `Concept` role is used to annotate fragments of text that represent a single unit of information in the domain. It can be a named entity, or a common noun, adjective or verb, that represents a concept relevant in the textual domain. Hence, almost every word or phrase that carries a singular meaning is annotated as `Concept` (or one of its derivatives, as explained next). Tokens such as articles, prepositions and conjunctions which only carry a grammatical function but not a semantic meaning are not annotated.

As an example, consider the sentence: "*El asma afecta las vías respiratorias*"[3]. In this sentence, the word *asma* is a clearly distinguishable concept in the health domain, whose meaning is independent of its grammatical role in the sentence. Some concepts such as *vías respiratorias* are multi-word, either because the single words that compose it are meaningless by themselves, or because the concept formed by their union is different from the individual meanings. In this case, even though *vías* and *respiratorias* by themselves have individual well-defined meanings, the concept *vías respiratorias* has a very definite meaning in the health domain that makes it a single unit of information, i.e., an specialist in the domain can clearly identify it.

3.2 Actions

An `Action` is a specific type of `Concept` which indicates a process or event, that some other concept can perform or receive the effects of, or an interaction between concepts. In the previous example, *afecta* is an action. An `Action` can be linked to relevant concepts by two semantic roles: `subject` and `target`. The `subject` is the concept that produces the action, while the `target` is the concept that receives the ef-

[3] In English: *Asthma affects the respiratory tract.*

fect of the action. In the previous example, the subject of *afecta* is *asma*, and the target is *vías respiratorias*. An Action can have zero or more subjects and/or targets. Figure 2 shows a graphical representation of the previous sentence with the corresponding Concepts and Actions, and the respective subject and target annotations.

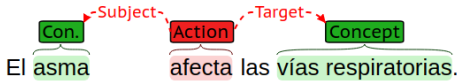

Figure 2: Annotation of Concepts and Actions in an example sentence.

In the previous example the Action is indicated by a word with the grammatical role of verb, which is intuitively the most common case. However, an action can also be indicated by a word with another grammatical role, such as nouns. For example, in the phrase "*...el empeoramiento de los síntomas...*"[4], the word *empeoramiento* is still considered an action even though it is not a verb, since it describes a process or event that happens to some other concept. Thus, the semantic role Action describes the intended meaning of a concept in the semantic domain, rather than its grammatical function in any specific sentence. If a domain concept expresses a process or event that produces effects on other Concepts, then it is an Action, even if it can be used in different grammatical functions.

3.3 References

A Reference is a type of Concept that has no specific semantic meaning, but it is necessary for grammatical reasons. It is used to annotate pronouns (e.g., *este*, *aquel*, etc.) and other referential elements when necessary, such as when they play the role of subject or target.

3.4 Predicates

A Predicate is used to form more complex concepts by combining, filtering or modifying other Concepts in a sentence. A common use case is for defining a subset of a Concept given some properties. For example, in the phrase "*...afecta a las personas mayores de 60 años...*"[5], the word *mayores* is annotated as a Predicate that filters some of the people. In conjunction with Predicates, any concept can play two additional roles: the domain or an argument of the predicate. In the previous example the domain is played with the Concept *personas*, and the only argument is *60 años*.

This construction gives rise to a new concept, that of people older than 60 years, which can be understood as the application of the filter *mayores* on a set of elements defined by the Concept *personas*, of whom those with the argument *60 años* are selected. The new complex concept built this way is represented in the sentence by the Predicate itself. Hence, to continue with the previous example, if we want these "older people" to play the target role then the corresponding annotation goes from the Action to the Predicate, as shown in Figure 3. It would be a mistake to say the subject of *afecta* is *personas* because this concept represents all people. Hence, the Predicate is used to represented not the filtering operation itself, but actually the filtered concept.

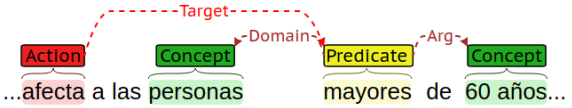

Figure 3: Annotation of Predicates and Actions in an example sentence.

3.5 Composing concepts

Just as Predicates can be used to define composite concepts, this can also be accomplished by considering an Action as the subject or target of another. For example, in the sentence "*Los empleados dedicados al cuidado de la salud están expuestos a riesgos laborales*"[6], there is complex concept involving *empleados*, *cuidado* and *salud*. This concept then acts as the target of *expuestos*, since it is not all employees that are exposed to hazards, but only those dedicated to health care (see Figure 4). This strategy can also be used to represent nominalizations, where the nominalized verb can be annotated as an Action and the corresponding subject and target construct the complex concept.

[4] In English: ... *the worsening of symptoms...*
[5] In English: ...*affects people older than 60 years...*
[6] In English: *Employees dedicated to health care are exposed to occupational hazards.*

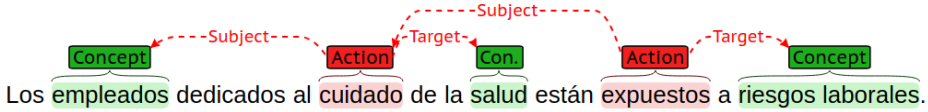

Figure 4: Annotation of composite concepts formed when an `Action` is `subject` of another.

3.6 Taxonomic relations

`Actions` and `Concepts` allow the capture of a large part of the semantic meaning of a sentence, by annotating as actions all the concepts that indicate any interaction between other concepts. However, some specific types of interactions are so common, that they are considered in many knowledge domains as building blocks for ontological of taxonomic representations. Such is the case of *hypernymy/hyonymy* pairs (i.e., `is-a` relations) and *meronym/holonym* pairs (e.g. *part-of* relations), which form the core of several knowledge bases.

These two types of relations are very common in most knowledge domains, and there many different textual variants to express these ideas. Arguably, it is better to explicitly represent them as relations between concepts, rather than resorting to annotate as an `Action` forms of the verb *to be*. Furthermore, an explicit annotation of these relations enables automatic knowledge discovery systems trained on these annotations to extract more compact and concise structures of knowledge, since there is no additional interpretation necessary.

The relations *is-a* and *part-of* can be explicitely indicated in the text by the appereance of common textual patterns (e.g., Hearst patterns (Hearst, 1992)). However, we also consider their annotation even when no explicit textual cues appear. For example, in the phrase "*...el corazón y otros órganos...*"[7] it is implicitly begin stated that *corazón* `is-a` *órganos*. A similar case is the example "*...el corazón y otras partes del cuerpo...*"[8] that implicitly indicates that the heart is a part of the body.

The relation `same-as` is used to indicate synonyms, or concepts that are considered equal in the document's domain. It can also be used when some simple concept is defined by describing it as another more complex concept, such as in the following example: "*Una ampolla es la piel que cubre una herida*"[9]. In this example, the `Concept` *ampolla* is being defined as another complex concept, formed by the `Action` *cubre* with `subject` *piel* and `target` *herida*. Hence, in this example the sentence is annotated as shown in Figure 5.

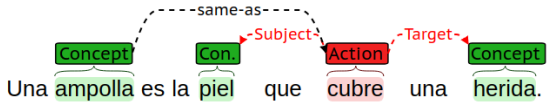

Figure 5: Annotation of a `same-as` relation in a definition.

The relation `has-property` is used to specify that a concept has a property or characteristic, or can be described by another concept. The simplest example is "*...el asma es peligrosa...*"[10], in which the `Concept` *asma* is related by `has-property` to the `Concept` *peligrosa*.

For all the taxonomic relations, we only consider the annotation when the sentence actually implies the existence of such relation, even if the implication is implicit. In no case we consider their annotation based solely on external or domain knowledge.

3.7 Causation and entailment

The previous 4 semantic relations are useful for capturing the taxonomic structure of the knowledge expressed in natural text. Two additional relations are defined for capturing logical connections between concepts: `causes` and `entails`. The relation `causes` is used to express that some event (identified in general as a `Concept`) is a possible cause for another event. An example is "*El asma provoca que las vías respiratorias se inflamen*"[11], annotated as shown in Figure 6. This relation indicates causation, not correlation or logical implication. Hence, it must be clearly stated in a sentence that there is a direct causation link between events. There is also a degree of uncertainty implied in the causation, which means that if *A* `causes` *B*, it doesn't necessarily imply that

[7] In English: *...the heart and other organs...*
[8] In English: *...the heart and other parts of the body...*
[9] In English: *A blister is the skin that covers a wound.*
[10] In English: *...asthma is dangerous...*
[11] In English: *Asthma causes the respiratory tract to become inflamed.*

every time *A* happens *B* will follow, or that anytime *B* happens, is due to *A*.

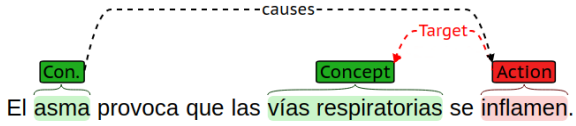

Figure 6: Annotation of the relation `causes`.

In contrast, the relation `entails` is used to denote a logical implication. In this case, it is not necessary for events to be related by causation at all; what must hold is that when some assertion *A* is true then it is always the case that assertion *B* is true. The annotation of causation and entailment avoids annotating several words and phrases that share the same semantic meaning. For example, in Figure 6 we refrain from annotating "*provoca*", since the actual meaning is already represented by `causes`.

3.8 Contextualization

Sometimes concepts only participate in certain relations with a precondition, such as during a specific period of time, in a specific location, or with some additional properties. An example is the sentence "*El dengue en estado avanzado es peligroso*"[12]. In this sentence the annotation *dengue* `has-property` *peligroso* fails to capture the whole semantic of the message, since dengue is not necessarily always dangerous (according to the sentence), but only in the specific situation when it is in advanced stage. For these situations, our model includes three contextual relations: `in-time`, `in-place` and the more general `in-context`. The previous sentence is annotated as shown in Figure 7.

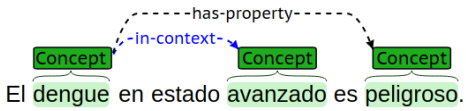

Figure 7: Annotation of the relation `in-context`.

The difference between contextual relations and the rest is that they do not define an assertion, but are only useful for building more complex concepts. For example, the annotation *dengue* `in-context` *avanzado* does not say that dengue always has the quality of being advanced. It is only when linked by `has-property` (or another relation) to other concepts, that this construction is meaningful. For this reason it is not correct to interchange `in-context` with `has-property`, since a `has-property` relation does state a specific assertion by its own.

3.9 Attributes

Four additional Boolean attributes can be attached to any concept to further qualify or describe it: `negated`, `uncertain`, `diminished` and `emphasized`. These attributes are used to avoid annotating stopwords such as *no*, *mucho*, *poco*, *puede*, and instead directly attaching the corresponding qualifier to the concept itself. These attributes also capture the intended negation, uncertainty or emphasis even when it is implied and not explicitly indicated by another word. An example is the phrase: "*...en ocasiones cura...*"[13] in which there is an implied uncertainty in the `Action` *cura*.

4 Annotation methodology

In this section, we briefly describe a methodology for creating a corpus based in this annotation model. At the moment of writing this process is being applied to the annotation of a corpus of 1000 Spanish sentences in the clinical text domain. This corpus is the main evaluation scenario for the eHealth-KD challenge to be hosted at IberLEF 2019[14]. The partial annotations and corpus statistics are available online[15].

The annotation process begins with the creation of a small collection of annotated sentences (i.e., a trial corpus) by a group of expert annotators. The selected sentences should cover all the important annotation patterns, and ideally, the most significant sources of ambiguity. From this trial corpus, an annotation guide can be constructed, that contains example annotations of all the semantic roles and relations defined. This guide defines the annotation protocol and also how to disambiguate conflicting patterns. The annotation guide is used as reference by the rest of the annotators during the whole process. For the annotation process we propose the following stages:

1. Manually tagging a set sentences independently by different non-biomedical experts.

[12] In English: *Dengue in advanced stage is dangerous.*

[13] In English: *...ocassionally it heals...*

[14] https://knowledge-learning.github.io/ehealthkd-2019

[15] https://github.com/knowledge-learning/satr-ann

Each sentence is tagged by two different annotators. Annotators are allowed to discuss general strategies, but should not discuss the specific sentences they are assigned. When in doubt, they will refer to the annotation guide and the trial examples.

2. Merging and normalization of tagging sentences between two annotators. In this case, another annotator selects the best annotations when contradictions exist. This stage can be aided by merging scripts that automatically detect and highlight conflicts.

3. The normalized sentences are verified and agreed upon by a committee of expert researchers in natural language processing, that decide which sentences are finally included in the corpus. Alternatively, if all members of the committee agree than a different annotation improves a specific sentence, it can be changed, but this situation should be the exception rather than the norm.

After the three stages, the set of manually annotated and revised sentences constitute the new corpus. These sentences should be then evaluated as described in Section 4.1.

4.1 Annotation evaluation

To evaluate the manual annotation agreement in the corpus, we propose to compute a micro-average of all the matches between every pair of annotations of the same sentences. This comparison can be performed in two stages. First, when the non-expert annotators label all the original sentences, each sentence receives annotations from two different people. Second, after the sentences are combined and revised by the expert committee, they can be compared to the original sentences, to understand how much the corpus changed between non-expert annotations in the review process. Since the annotation task involves selecting subsets of text and labelling them with different tags, we propose to use an F_1 metric (as opposed to the most common Kappa metric), such as the one used by Moreno et al. (2017) for the DrugSemantics corpus. Since the annotation involves fragments of text, it is important to consider partial agreement between annotators. For this purpose, we propose to score partially matching spans of texts proportionally to the length of their intersection.

Another important evaluation metric is the human performance in this task, since corpora created with this annotation models are frequently used for machine learning tasks. We propose that after the corpus is built an additional annotator performs a manual labelling of a predefined subset of the sentences. This annotator can be trained with the same annotation guidelines, but should not have been exposed to this specific subset of sentences before. This can be used as a baseline for human performance and can be compared to the performance of different algorithms trained in the corpus. In the eHealth-KD challenge, this strategy will be applied to provide a human performance metric for comparative purposes.

4.2 Annotation guidelines

The most relevant characteristic of the annotation model presented in this research is that it intends to represent the semantic, rather than the syntax of sentences. For this purpose, it is necessary to avoid in annotators incorrect mindsets that fix semantic roles to grammatical functions (e.g., considering that verbs are almost always actions). The correct process is understanding the semantic meaning of a sentence first, and then representing it using the annotations. A useful heuristic is to attempt to reconstruct a sentence from the annotations, possibly with a different wording than the original, but with the same meaning. It is also important to annotate all the concepts that appear in the sentence even if they cannot be eventually interrelated. Finally, we prefer annotating the most explicit relation possible; for example, using `cause` instead of using an Action such as "*produce*" or "*provoca*", if `cause` accurately captures the semantic meaning of the corresponding phrase.

4.3 Annotation tools

The tool proposed for all the manual annotation process is BRAT (Stenetorp et al., 2012). This tool makes it possible to visually select text portions, assign labels and connect them by relations, through a simple web interface that requires little to no previous training. Even though BRAT allows a limited form of collaborative annotation, we actually prefer that different annotators work in different copies of the text (Stage 1), and afterwards perform an automatic merging process using custom scripts that output a BRAT-compatible result. Then, in Stage 2, the expert who performs the normalization can continue to use BRAT to correct

mistakes. Furthermore, the web interface of BRAT enables online collaboration between annotators that are not physically close. For our model, we provide relevant configuration files for BRAT and 50 annotated examples sentences online[16].

5 Conclusions and future work

This research proposes a general-purpose annotation model that captures a broad range of semantic information from textual content, based on Subject-Action-Target triplets plus additional semantic relations. This model extends the annotation model used by the eHealth-KD corpus, with the addition of two semantic roles (`Predicate` and `Reference`), the representation of causation and entailment, and the possibility of identifying contextual qualifiers. Theses additions allow capturing more complex semantic information than the previous model. Our ongoing efforts focus on annotating a large corpus of clinical text in Spanish for supporting shared evaluation campaigns.

The semantic roles and relations defined map to common concepts and relations used in knowledge bases and ontologies, which simplifies the task of building semantic networks from the annotated text. In the future we will focus on this mapping stage, which will also require linking these concepts to entities hosted at shared knowledge bases, such as DBpedia (Auer et al., 2007) and UMLS (Bodenreider, 2004). In addition, we also plan to pursue the annotation of clinical text, and extending to additional languages and other domains, such as news, scientific papers, encyclopedic articles and others.

Acknowledgments

Funding: This research has been supported by a Carolina Foundation grant in agreement with 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; SocialUniv 2.0 (ENCARGO-INTERNOOMNI-1); and PINGVALUE3-18Y.

The authors would like to thank the team of annotators from the School of Math and Computer Science, at the University of Havana.

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

[16] https://github.com/knowledge-learning/satr-ann/tree/master/data/v1

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

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

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1306–1313. AAAI Press.

Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. *KR-MED 2006*, page 87.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

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

A Intxaurrondo, JC de la Torre, H Rodriguez Betanco, M Marimon, JA Lopez-Martin, A Gonzalez-Agirre, J Santamarıa, M Villegas, and M Krallinger. 2018.

Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of spanish clinical abbreviations: the barr2 corpus. SEPLN.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision – ECCV 2016*, pages 852–869. Springer International Publishing.

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. 2018. Overview of tass 2018: Opinions, health and emotions.

Jonathan May and Jay Priyadarshi. 2017. 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.

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

Martha Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the generative lexicon conference*, pages 9–15. GenLex-09, Pisa, Italy.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using abstract meaning representation. *BioNLP 2017*, pages 126–135.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966.

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.

Matthew S. Simpson and Dina Demner-Fushman. 2012. *Biomedical Text Mining: A Survey of Recent Progress*, pages 465–517. Springer US, Boston, MA.

Frederic Stahl, Bogdan Gabrys, Mohamed Medhat Gaber, and Monika Berendsen. An overview of interactive visual data mining techniques for knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):239–256.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. 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*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.