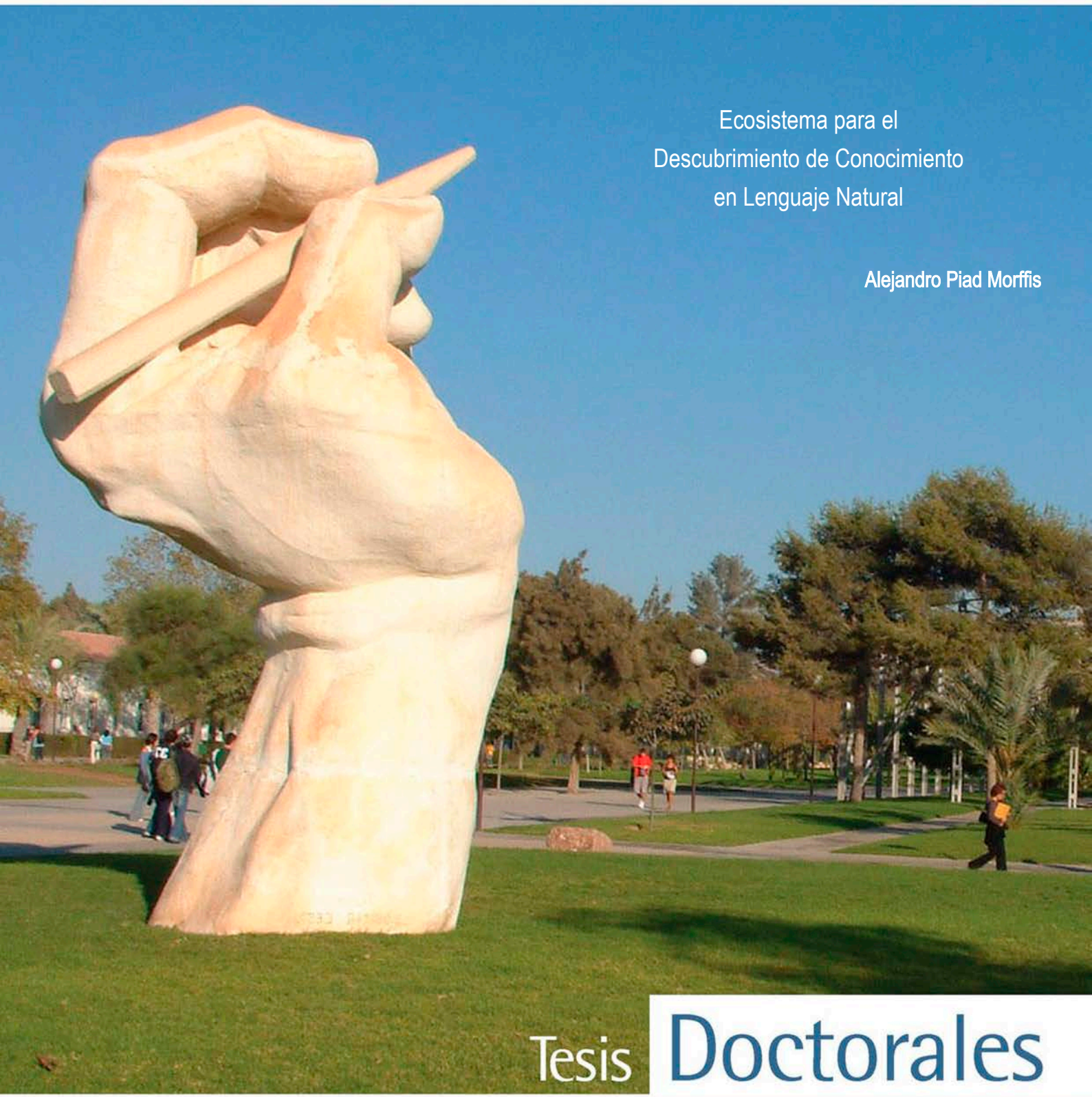
Universitat d'Alacant
Universidad de Alicante

Ecosistema para el Descubrimiento de Conocimiento en Lenguaje Natural

Alejandro Piad Morffis



Tesis Doctorales
UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Ecosistema para el Descubrimiento de Conocimiento en Lenguaje Natural

Alejandro Piad Morffis

Tesis presentada para aspirar al grado de

DOCTOR POR LA UNIVERSIDAD DE ALICANTE

DOCTORADO EN INFORMÁTICA

Dirigida por

Dr. Yoan Gutierrez Vazquez
Dr. Yudivian Almeida Cruz
Dr. Rafael Muñoz Guillena

Esta tesis ha sido co-dirigida y financiada por
la Universidad de Alicante y la Universidad de La Habana.

TESIS DOCTORAL EN FORMA DE COMPENDIO DE PUBLICACIONES

Ecosistema para el Descubrimiento de Conocimiento en Lenguaje Natural

El presente documento contiene una síntesis del trabajo realizado por Alejandro Piad Morffis, bajo la dirección de Dr. Yoan Gutierrez Vazquez, Dr. Rafael Muñoz Guillena y Dr. Yudivián Almeida Cruz, para optar por el grado de Doctor en Informática. Se presenta en la Universidad de Alicante y se estructura según la normativa establecida para la presentación de tesis doctorales en forma de compendio de publicaciones: una primera parte con una síntesis, una segunda parte que reproduce las publicaciones científicas realizadas y una tercera parte con las conclusiones.

A mi familia y amigos, por hacerme quien soy.

A Sue, por enseñarme a querer más.

A papá, por creer.

Agradecimientos

Cuando empecé la aventura que me trajo al día de hoy, allá por el año 2008, yo tenía grandes expectativas pero ningún plan concreto. Sabía que quería estudiar en la Universidad de La Habana, y que quería hacer algo con computadoras, pero nada más. Mi padre se sentó conmigo, él al timón, yo en el asiento del acompañante, como siempre hacíamos por esa época, esperando a que mi hermana, quien entonces estudiaba violín, terminara sus clases.

Papá me preguntó cuáles eran mis aspiraciones ahora que comenzaba la Universidad. Él siempre recordaba la Universidad como la época de su vida donde más grande se hizo su mundo, y creo que por eso quería que la mía fuera una experiencia excepcional. Yo le dije simplemente que quería ser el mejor. El mejor de mi clase, el mejor de toda la carrera, vaya, el mejor del mundo incluso. Y en vez de decirme cuan ridículas eran esas expectativas, me dijo "pues hagamos un plan para lograr eso".

Y así lo hicimos, planeamos allí en media hora los pasos que supuestamente me iban a llevar a cumplir esas ridículas metas. Pasos que tenían que ver con dedicar no se cuántas horas a estudiar, con hacerme alumno ayudante lo antes posible, con trabajar en proyectos adicionales. Hoy sé que no había ninguna garantía de que esos pasos me llevaran a ser el mejor del mundo, pero al menos me trajeron hasta aquí. Y así aprendí que los sueños no están para cumplirse, están para empujarte a llegar más lejos, mucho más lejos de lo que tu pudieras imaginar, aunque muchas veces en direcciones totalmente diferentes de las que soñaste.

Este sueño lo he compartido con muchísimas personas, tantas que me es imposible agradecerles a todos por las huellas que han dejado en mi vida. Trataré de mencionar a algunos de los imprescindibles, pero pido disculpas de antemano por todos aquellos que seguro me faltarán.

En primer lugar, tengo que agradecer infinitamente a mi padre. Agradecerle por muchas cosas simples y mundanas, como enseñarme a montar bicicleta, a nadar, a disfrutar de los libros, a conducir, a comer saludable, a querer a los demás por sus defectos y no a pesar de ellos, y a soñar en grande. Pero lo más importante que me dió mi padre no fueron ni habilidades específicas ni su filosofía de vida. Papá siempre creyó que yo podía lograr lo que me propusiera, incluso cuando eso no coincidiera con lo que él veía más importante o valioso, incluso aunque que mis planes y expectativas me llevaran lejos de las suyas. Y no es cierto, uno no puede lograr todo lo que se propone, claro que no. Su propia vida estuvo llena de metas fallidas y expectativas frustradas, como la de todos. Pero si uno no cree que puede, no tiene sentido ni siquiera intentarlo. Y por eso el superpoder de mi padre era creer, creer que todo era posible, que solo dependía de uno mismo, y que no había obstáculo suficientemente grande.

Si bien mi padre me dio gran parte del combustible que quemé para hacer lo que he hecho, nunca lo hubiera logrado sin mi madre. Cuando las cosas se ponían de verdad complicadas, cuando parecía que no había más solución que rendirse, mi madre siempre encontraba una salida, una forma de reinventarse para seguir adelante. Lo hizo cuando yo era un niño y no alcanzaba la comida para los tres, lo hizo de nuevo cuando nació mi hermana, lo volvió a hacer años después cuando todo lo que tenían construido tambaleó, y después de aquel octubre de 2019 cuando la vida de todos nosotros dio un vuelco repentino y definitivo. De mi madre aprendí muchas cosas, pero quizás la más importante es esa: que uno tiene el poder de reinventarse tantas veces como sea necesario, por uno mismo, y por los que uno quiere. Y que nadie más que tú mismo puede realmente derrotarte.

Mi hermana ha sido la tercera constante en mi vida, o al menos la mayor parte de ella. Alenia trajo una música especial a mi mundo, una forma de ver la vida, de querer a los demás, y de simplemente ser, que complementa de manera perfecta mi propia forma de ver el mundo. Hay una cosa en especial que compartimos, una admiración mutua, que nos hace estar cerca aunque haya distancia física. Siendo yo el mayor, siempre he sentido que mi hermana es mi fan número uno. Lo que ella no supone, porque casi nunca lo he dicho, es que yo también soy su fan número uno, y que a pesar de los años de supuesta ventaja que le llevo, son muchas las enseñanzas que he sacado de verla crecer, convertirse en una persona capaz y feliz de cometer sus propios

errores, bajo sus propias condiciones, siendo una fuente de energía vital para muchos a su alrededor, y sin perder su propia identidad en el proceso. Y junto a Alenia, tengo que agradecer a Andy por acompañarla, por complementarla, y por traer a nuestra familia su júbilo, sus ganas de vivir, de reír, y de cantar. Ellos dos me han enseñado el valor de perseguir tus sueños, contra viento y marea, a pesar de todas las opiniones que puedas tener en tu contra.

Mi familia no es muy grande, y no siempre han podido estar todos los que hubiesen querido. Pero de cada uno he aprendido algo, y a cada uno le dedico también un pedacito de estas páginas. A mi abuela Sofía, que estuvo ahí desde mis primeros días de escuela, y que sigue ahí a pesar de los pesares. A mi abuelo Roberto, que fue un ejemplo de honradez, de trabajo duro, y de entrega por sus seres queridos, y quien desgraciadamente no pudo ver a sus nietos crecer tanto como hubiese querido, pero estoy seguro estaría orgulloso. A mi abuelo Pepín, que a pesar de lo poco que pudo compartir conmigo, guardo recuerdos hermosos. A mis tíos y primos por ambas partes de la familia, que han estado en momentos buenos, y algunos en otros momentos no tan buenos. A mi tío Frank en especial por el cariño incondicional, y a mi primo Francito, como siempre le llamaré, por tantas horas que compartimos de niños, y las que seguro nos quedan por compartir.

Quiero agradecer también a mi segunda familia, la que no me tocaba, pero me acogió como si fuera un hijo más. A mis suegros, Estela y Armando, que me enseñaron a ver la vida de forma un poco diferente; sus enseñanzas sí que me han marcado, mucho más de lo imaginan, y mucha buena culpa tienen de la persona que soy hoy. A Wendy, que ha sido una hermana más, por su forma de sentir y querer que no conoce fronteras ni limitaciones; y a Daniel, por su amistad, y por darme algunas de las horas de conversación más interesantes que he tenido. A mis abuelos adoptivos, Mimi, Gregorio, Ondina, y Armando, por su cariño immerecido; a Tony, Betty, Alain y Yahima, por hacerme parte de su familia. Y a todos los demás, tíos, primos, y abuelos, por acogerme como uno más en su familia gigante.

Amigos tengo muchísimos, algunos desde siempre, otros desde hace poco, y todos han dejado una huella en mi. En especial quiero agradecir a mis amigos del alma, con los que viví tres años compartiéndolo todo, desde el agua y la comida hasta las aspiraciones, los sueños, y las ganas de vivir. A Silvio, Charly, Ian Pedro, Pepito, y Luis Alberto, no tengo forma de agradecerles

por tantos años de estar ahí, en las buenas y en las malas, aunque estemos desperdigados por todos los continentes. A mis amigos y compañeros de la UH, con los que he aprendido no solo habilidades y conocimientos, sino formas nuevas de pensar y de ver la vida. A mis amigos de Alicante, de España, y del resto del mundo, que me han abierto los horizontes y me han dado nuevas esperanzas y expectativas. De todos ellos me llevo un trocito, y espero que todos tengan en cambio algo de mí. Y quiero agradecer también a los más chiquitos, que fueron mis estudiantes y ahora se han convertido en compañeros, a Jonpi, Roci, Daniel, Hian, Sadán, Estevanell, y todos los que vienen en camino. Ellos son mi mayor fuente de orgullo, y por cada cosa que hayan podido aprender de mi, hay algo que he aprendido también yo de ellos.

Desde mis primeras letras hasta el día de hoy, todo lo que pueda haber logrado lo debo en gran medida a mis profesores, que me enseñaron a leer, a sumar, a derivar, a programar, a escribir artículos, a hablar en público, a dirigir un proyecto. Son demasiados para nombrarlos a todos, pero sepan que los recuerdo, y si escogí esta profesión, que considero la más bella, es por la admiración que me han hecho sentir de su trabajo. En especial, quiero agradecer a Marrero, por inspirarme su amor por la ciencia, y enseñarme la belleza que hay en entender el universo. A Lupi, por prestarme atención en aquel primer año, y empujarme siempre a saber más. A Bolu, por enseñarme a pensar como un científico, y animarme a seguir ese camino. A Luciano y Katrib, por inspirarme con su energía inagotable y demostrarme que no hay cansancio que valga si uno ama su trabajo. Por supuesto, a mis tutores de Alicante, Yoan, Rafa y Andrés, por confiar en mi, y guiarme por este viaje con infinita paciencia. Y muy especialmente, a Yudy, por ser maestro y amigo, por enseñarme que siempre podía aspirar a más incluso aunque el resto del mundo no estuviera de acuerdo, y por estar siempre dispuesto a compartir este sueño.

Finalmente, quiero agradecer a mi esposa, mi mejor amiga, y mi inseparable compañera de equipo, Suilan. Ella me ha enseñado demasiadas cosas para resumirlas aquí, pero la lección más importante que he aprendido es que, en última instancia, nadie más que tú puede decirte que algo es imposible. Suilan expandió mi universo rompiendo límites que yo ni me daba cuenta que existían, y se atrevió a caminar por ese universo conmigo, compartiendo miles de horas de frustración, pero también innumerables alegrías. Ella es

mi mayor crítico, mi más valioso consejero, y el más importante miembro en todos mis equipos. Nada de lo que pueda decir aquí sería suficiente para resaltar la importancia que ha tenido su presencia en mi vida, como una gigante roja cuya fuerza gravitacional cambia todo a años luz de distancia. De no haberla conocido, no tengo idea de dónde estaría, pero seguro seríamos ambos muy diferentes a cómo somos. Y por si fuera poco, me ha dado la alegría mayor de saber que pronto seremos padres, y tendremos así una nueva aventura, seguramente más difícil que todas las anteriores, pero una aventura que también haremos juntos.

Cuando empecé este viaje hace más de 13 años, nunca imaginé donde estaría hoy. Imaginé muchas cosas, algunas que pasaron, y muchas otras que no pasaron. Por el camino he aprendido tantas cosas, cosas sobre el mundo, y cosas sobre mí, que no estoy seguro de pueda decir que aquella persona que empezó con aquel plan en aquella tarde de 2008 sea la misma que escribe estas letras. Muchas puertas se me abrieron, gracias a muchas personas que no he alcanzado a mencionar, así como otras puertas se cerraron. He cruzado aquellas que consideré mejor, y aunque estoy seguro que muchas veces dejé ir oportunidades que otros veían valiosas, si tuviera la posibilidad de repetir cada una de mis decisiones trascendentales, haría exactamente lo mismo sin pensarlo dos veces. Cualquier error que haya cometido es total responsabilidad mía, y asumo la culpa con gusto, porque son esos errores los que me han traído a poder escribir estas palabras hoy. Cualquier ventaja, justa o injusta que haya tenido, se la debo a alguien, y por esas estoy agradecido.

Cierro con este agradecimiento general a las oportunidades que me ha dado la vida, este sueño de convertirme en un experto en el campo de estudio que escogí ejercer, y de ganarme el respeto y admiración de mis pares. De ahora en adelante, tengo muchos sueños nuevos. Todavía sueño con ser el mejor: el mejor padre, el mejor maestro, el mejor amigo. Y lo sé, los sueños no están ahí para cumplirse, están para empujarte a llegar más lejos, mucho más lejos de lo que tu pudieras imaginar.

A todos los que han estado y los que siguen estando dispuestos a perseguir estos sueños conmigo, muchas gracias.

Alejandro Piad Morffis
La Habana, a 15 de Octubre de 2021

Esta investigación ha sido desarrollada de forma conjunta en la Universidad de Alicante (España) y la Universidad de La Habana (Cuba), entre noviembre de 2017 y septiembre de 2020, en sucesivas estancias de investigación cofinanciadas por ambas instituciones. Por la Universidad de Alicante, el Departamento de Lenguajes y Sistemas Informáticos ha soportado esta investigación a través de los proyectos SIIA (`PROMETEO/2018/089`, `PROMETEU/2018/089`) y LIVING-LANG (`RTI2018-094653-B-C22`). Por la Universidad de La Habana, la Facultad de Matemática y Computación y el Departamento de Inteligencia Artificial y Sistemas Computacionales han soportado esta investigación.

Resumen

La creciente cantidad de información publicada en línea presenta un reto significativo para la comunidad científica. La disponibilidad de estos recursos permite acelerar las investigaciones en múltiples ramas de la ciencia, al conectar resultados de diferentes grupos de investigadores. Sin embargo, el volumen de información producido es imposible de procesar por humanos en su totalidad, por lo que la comunidad científica desperdicia tiempo y recursos en redescubrir los mismos resultados, debido a la falta de comunicación. La aplicación de técnicas de inteligencia artificial permite construir sistemas computacionales que ayuden a los investigadores a buscar, analizar y conectar la información existente en grandes volúmenes de datos. Este proceso se denomina descubrimiento automático de conocimiento y es una rama de investigación con un creciente interés.

El dominio de la salud es uno de los escenarios en los que el descubrimiento de conocimiento automático puede producir un mayor impacto en beneficio de la sociedad. La reciente pandemia de COVID-19 es un ejemplo donde la producción de artículos científicos ha superado con creces la capacidad de la comunidad científica para asimilarlos. Para mitigar este fenómenos se han publicado recursos lingüísticos que permitan construir sistemas de descubrimiento automático de conocimiento. Sin embargo, el descubrimiento de conocimiento requiere no solo de recursos lingüísticos, sino que necesita recursos computacionales e infraestructura disponibles para evaluar los resultados sistemáticamente y comparar objetivamente enfoques alternativos.

Este trabajo describe un ecosistema que facilita la investigación y el desarrollo en el descubrimiento de conocimiento en el dominio biomédico, específicamente en idioma español, aunque puede ser extendido a otros dominios e idiomas. Con este fin, se desarrollan y comparten varios recursos

con la comunidad investigadora, incluido un nuevo modelo de anotación semántica, cuatro corpus con más de 3000 oraciones y 40,000 anotaciones semánticas realizadas manualmente, así como recursos computacionales para construir y evaluar técnicas de descubrimiento automático de conocimiento. Entre estos recursos se ofrecen implementaciones *baseline* de algoritmos de descubrimiento de conocimiento que sirvan de base para construir soluciones más avanzadas. Además, se define una tarea de investigación con criterios de evaluación objetivos y se configura y mantiene un entorno de evaluación en línea que permite a los investigadores interesados en esta tarea obtener retroalimentación inmediata y comparar sus resultados con el estado del arte. Como caso de estudio, se analizan los resultados de varios equipos de investigadores en cuatro ediciones consecutivas de un desafío competitivo organizado en base a estos recursos.

A partir de las experiencias obtenidas durante el proceso de anotación manual se diseña una estrategia de anotación asistida que permite reducir considerablemente el tiempo de anotación humano. El enfoque ayuda a los anotadores humanos seleccionando inteligentemente las oraciones más informativas para anotar y luego pre-anotarlas con algunas entidades y relaciones semánticas altamente precisas. Esta estrategia se evalúa en los corpus desarrollados en esta investigación, y se publica en forma de una herramienta computacional disponible para la comunidad científica.

El ecosistema construido proporciona un entorno de aprendizaje y evaluación eficaz para fomentar la investigación en el descubrimiento de conocimientos tanto en documentos de contenido biomédico como en otros dominios. Los corpus anotados pueden ser utilizados para entrenar y evaluar sistemas computacionales de descubrimiento de conocimiento, y compararse con el estado del arte de forma automática. Así mismo, las herramientas computacionales desarrolladas pueden servir para construir nuevos sistemas y para crear nuevos recursos lingüísticos en otros idiomas o dominios. Todos los recursos desarrollados en esta investigación están disponibles públicamente para su uso por la comunidad científica[1].

[1] https://ehealthkd.github.io

Abstract

The increasing amount of information published online presents a significant challenge for the scientific community. The availability of these resources makes it possible to accelerate research in multiple branches of science, by connecting results from different research groups. However, the volume of information produced is impossible to process by humans in its entirety, hence, the scientific community wastes time and resources in rediscovering the same results, due to lack of communication. The application of artificial intelligence techniques allows the construction of computer systems that help researchers to search, analyse and connect the existing information in large volumes of data. This process is called automatic knowledge discovery and it is a branch of research with increasing interest.

The ehealth domain is one of the scenarios in which automatic knowledge discovery can produce a greater impact for the benefit of society. The recent COVID-19 pandemic is an example where the production of scientific articles has far exceeded the capacity of the scientific community to assimilate them. To mitigate this phenomenon, linguistic resources have been published, useful for building automatic knowledge discovery systems. However, knowledge discovery requires not only linguistic resources, it needs computational resources and available infrastructure to systematically evaluate results and objectively compare alternative approaches.

This work describes an ecosystem that facilitates research and development in automatic knowledge discovery in the biomedical domain, specifically in the Spanish language, although it can be extended to other domains and languages. To this end, several resources are developed and shared with the research community, including a new semantic annotation model, four corpora with more than 3,000 sentences, and 40,000 manually performed semantic annotations, as well as computational resources to build and eva-

luate techniques for automatic knowledge discovery. These resources include baseline implementations of knowledge discovery algorithms that serve as the basis for building more advanced solutions. In addition, a research task is defined with objective evaluation criteria and an online evaluation environment is configured and maintained that allows researchers interested in this task to obtain immediate feedback and compare their results with the state of the art. As a case study, the results of several teams of researchers in four consecutive editions of a competitive challenge organised based on these resources are analysed.

Based on the experiences obtained during the manual annotation process, an assisted annotation strategy is designed that produces a considerable reduction in human annotation time. The approach helps human annotators by intelligently selecting the most informative sentences to annotate and then pre-annotating them with some highly accurate semantic entities and relationships. This strategy is evaluated in the corpus developed in this research, and is published in the form of a computational tool available to the scientific community.

This computational ecosystem provides an effective learning and assessment environment to foster research in knowledge discovery in both biomedical documents and other domains. The annotated corpus can be used to train and evaluate computational knowledge discovery systems, and compare them automatically with the state of the art. Likewise, the computational tools developed can be used to build new systems and to create new linguistic resources in other languages or domains. All the resources developed in this research are publicly available for use by the scientific community[2].

[2] https://ehealthkd.github.io

Índice general

Índice de figuras

Índice de tablas

Parte I

Síntesis de la Tesis

Capítulo 1

Introducción

El crecimiento exponencial de Internet en las últimas décadas ha producido un excedente masivo de información textual en todas las áreas del desarrollo humano. Este escenario presenta tanto una oportunidad como un desafío para los investigadores. Por un lado, conectar resultados publicados en diferentes fuentes permitiría potencialmente descubrir nuevo conocimiento que está actualmente disperso. Por otro lado, la totalidad de la información disponible no puede ser procesada manualmente en un plazo razonable. Por lo tanto, los esfuerzos se han dirigido recientemente hacia el diseño de técnicas automáticas que pueden descubrir información relevante en grandes corpus, hacer conexiones lógicas y sintetizar conocimientos útiles. Este proceso se denomina descubrimiento automático de conocimiento y es una rama de investigación con un creciente interés [2]. El primer paso en muchas de estas técnicas implica la recopilación, el procesamiento y la anotación de datos que se pueden utilizar para entrenar algoritmos de aprendizaje automático o construir sistemas expertos mediante el uso de técnicas de procesamiento de lenguaje natural.

El dominio de la salud digital es de gran interés para la comunidad investigadora dados los beneficios sociales potenciales derivados de la aplicación de tecnologías automáticas de descubrimiento de conocimiento. La comunidad científica ha producido abundantes corpus anotados en diferentes sub-dominios de este sector, desde conocimiento específico (p.e., interacciones entre medicamentos y enfermedades [3] o genes y proteínas [4]) hasta modelos

más generales independientes (p.e., informes de ensayos clínicos [5]). Los corpus y las tecnologías específicas del dominio son de importancia crítica en la medicina de alta precisión. Sin embargo, los sistemas creados para dominios muy específicos son más difíciles de generalizar y ampliar que los sistemas basados en conceptualizaciones de propósito más amplio. Por tanto, existe un interés creciente en el diseño de modelos de anotación y corpus con una semántica de propósito general que puedan usarse en una variedad de dominios o como una componente en sistemas especializados.

Además del dominio, el idioma es otra dimensión que ha sido foco de investigaciones recientes. La mayoría de los recursos lingüísticos más utilizados se basan en fuentes en idiomas inglés, motivados en parte por la abundancia de contenido disponible (p.e., enciclopedias en línea o artículos científicos), lo cual no es sorprendente dado que el inglés es el idioma predominante en la ciencia, la tecnología y las comunicaciones a nivel internacional. Sin embargo, los recursos basados en inglés a menudo no son directamente aplicables a otros idiomas. Aunque la traducción automática ha alcanzado una precisión impresionante en los dominios abiertos, sigue siendo un desafío crear recursos en varios idiomas, como el español, que son menos utilizados en dominios técnicos [6]. En lugar de centrarse en idiomas específicos, una línea de investigación alternativa es diseñar recursos que sean agnósticos al idioma basándose en características comunes. Esto permitiría su generalización a múltiples idiomas con poco esfuerzo.

1.1. Motivación

El diseño de modelos de anotaciones que pueden generalizarse a múltiples dominios e idiomas requiere una representación básica del lenguaje que cubra una amplia gama semántica. Además, estas representaciones deben ser tan independientes de la sintaxis y las reglas gramaticales como sea posible. El trabajo reciente de Estevez-Velarde et al. [7] sugiere que las tripletas Sujeto-Acción-Objeto pueden usarse para detectar una amplia cantidad de interacciones semánticas en lenguaje natural, independientes del dominio y relativamente independientes del idioma, ya que más del 75 % de los idiomas humanos emplean alguna variación de la estructura gramatical Sujeto-Verbo-Objeto [8]. Del mismo modo, varias representaciones ontológicas a menudo coinciden en una serie de relaciones de propósito

general, (por ejemplo, hipónimos—*is-a*—, holónimos—*part-of*—) que son útiles en cualquier dominio. Otras conceptualizaciones permiten capturar la semántica, centrada en la sintaxis, más cerca del lenguaje natural, como AMR (*Abstract Meaning Representation*) [9]. La construcción de corpus anotados con estructuras semánticas de propósito general como Sujeto-Acción-Objeto y relaciones ontológicas de alto nivel es el primer paso en el diseño de sistemas que pueden descubrir conocimiento automáticamente en una variedad de dominios y escenarios.

El descubrimiento automático de conocimiento requiere no solo recursos lingüísticos (por ejemplo, corpus anotados) sino también recursos e infraestructuras computacionales que permiten a los investigadores evaluar sistemáticamente sus resultados y compararlos objetivamente con enfoques alternativos. Esto requiere la definición formal de tareas y el diseño de métricas de evaluación objetivas que garanticen una comparación justa. Aún mejor es un entorno de evaluación disponible para la comunidad donde los investigadores puedan enviar sus resultados, garantizando que se apliquen los mismos criterios de evaluación y liberando a los investigadores de reproducir el entorno de evaluación. Dicho sistema también garantizaría un proceso de investigación más transparente y reproducible, y proporcionaría un repositorio centralizado de los enfoques existentes, ayudando a los nuevos investigadores a actualizarse sobre el estado del arte.

1.2. Trabajos Relacionados

En investigaciones recientes se han establecido diferentes relaciones semánticas para capturar el conocimiento en lenguaje natural, muchas de las cuales dan lugar a la construcción de corpus. En esta sección se presenta un breve resumen al estado del arte relacionado con la presente Tesis. En el Capítulo 2 se presenta un análisis detallado del estado de la técnica donde se describen los trabajos relacionados en mayor profundidad.

Esta investigación se centra tanto en corpus o modelos de anotación para representar el conocimiento en múltiples dominios, como aquellos específicamente diseñado para el dominio de la salud. La tabla 1.1 presenta siete características más relevantes para la propuesta planteada en esta investigación e indica cuáles de ellas están presentes en una muestra de corpus

del estado del arte. Se incluyen en la comparación las cuatro ediciones de un corpus desarrollado en el marco de esta investigación (*eHealth-KD*). Las características consideradas son:

1. *independiente del dominio:* aplicabilidad del esquema de anotación subyacente a cualquier dominio;
2. *independiente de la sintaxis:* capturar aspectos semánticos en lugar de relaciones sintácticas en oraciones;
3. *conocimiento ontológico:* soportar la herencia y la composición de conceptos;
4. *conceptos compuestos:* permitir la anotación de conceptos que involucran otros subconceptos;
5. *atributos:* utilizar atributos como cuantificadores (por ejemplo, número de ocurrencias) o calificadores (por ejemplo, grado de certeza);
6. *relaciones contextuales:* permitir relaciones que solo ocurren cuando están condicionadas por un contexto específico; y,
7. *causalidad/implicación:* incluir relaciones para representar causalidad y/o vinculación.

Los modelos de anotación de propósito general a menudo se usan en corpus extraídos de fuentes enciclopédicas, como *YAGO* [14] y *ConceptNet* [15], los cuales contienen hechos seleccionados automáticamente de Wikipedia (entre otras fuentes). Por el contrario, los modelos de anotación de dominio específico generalmente se emplean cuando la fuente está más restringida. Los ejemplos incluyen *Ixa MedGS* [10], que contiene conceptos relacionados con la salud para enfermedades, causas y medicamentos; *DrugSemantics* [11], que anota entidades sanitarias, medicamentos y procedimientos; y *DDI* [12], que anota las interacciones farmacológicas. Un término medio es el corpus *Bio AMR* [13], que aplica un modelo de anotación de propósito general (*Abstract Meaning Representation*, AMR) [9] a los documentos de salud. El corpus *eHealth-KD* es similar a este último en este sentido, ya que el modelo de anotación definido es general, pero se aplica específicamente a las oraciones de salud en esta investigación.

	Características	**Ixa MedGS** [10]	**DrugSemantics** [11]	**DDI** [12]	**Bio AMR** [13]	**YAGO** [14]	**ConceptNet** [15]	**eHealth-KD 2018**	**eHealth-KD 2019/20/21**
1	independiente del dominio				✓	✓	✓	✓	✓
2	independiente de la sintaxis	✓	✓	✓		✓	✓	✓	✓
3	conocimiento ontológico				✓	✓	✓	✓	✓
4	conceptos compuestos				✓			✓	✓
5	atributos		✓		✓	✓		✓	✓
6	relaciones contextuales				✓				✓
7	causalidad/implicación	✓			✓		✓		✓

Tabla 1.1: Comparación entre los corpus de *eHealth-KD* 2018 al 2021, con otros recursos similares con respecto a las características que definen la propuesta presentada.

La mayoría de los recursos antes mencionados se centran en capturar la semántica de las oraciones, en el sentido de que es probable que oraciones diferentes con los mismos hechos se anoten de manera similar. Se puede considerar que *BioAMR* es más dependiente de la sintaxis porque, aunque AMR es un modelo de anotación semántica —más abstracto, por ejemplo, que el análisis de dependencia— todavía depende en gran medida de la estructura gramatical de las oraciones. Por lo tanto, es probable que un cambio significativo en la estructura de la oración cambie la anotación, incluso si el mensaje semántico subyacente es el mismo. Por ejemplo, dado que AMR usa los roles de PropBank [16], cambiar una palabra por otra semánticamente similar, incluido un sinónimo, probablemente cambiará la anotación correspondiente y, por lo tanto, los roles disponibles. Esto también hace que AMR y recursos similares dependan del idioma, no solo en la práctica dada su dependencia de la existencia de bancos de palabras, pero también desde su definición. Al intentar aplicar AMR en español, Migueles-

Abraira et al. [17] muestra que aunque es teóricamente independiente del idioma, las guías de anotación existentes están sesgadas hacia el inglés y deben adaptarse para capturar otros fenómenos lingüísticos que no existen en este idioma. En esta investigación se intenta lograr un mayor nivel de independencia sintáctica, en parte mediante el uso de un conjunto más pequeño de entidades, relaciones y roles que AMR.

Una estrategia a menudo utilizada para alentar la investigación sobre una tarea específica es la organización de campañas de evaluación competitivas. En contraste con la investigación regular, estas campañas a menudo tienen una duración predefinida y los recursos de evaluación no se divulgan completamente (por ejemplo, las anotaciones para los conjuntos de prueba) para permitir una comparación justa en un entorno competitivo amigable. Uno de los más importantes ejemplos en el dominio de la salud es el *CLEF eHealth Evaluation Lab*, que ha propuesto numerosos eventos competitivos en el dominio médico, incluyendo tareas de reconocimiento de entidades nombradas [18] y extracción de información [19] en inglés, y en ediciones posteriores también en documentos en francés [20, 21]. En otras ediciones los informes médicos de MEDLINE, EMEA y fuentes similares se han anotado con trastornos, términos médicos, siglas y abreviaturas, que proporcionan escenarios de evaluación para varias tareas de procesamiento de lenguaje natural, incluyendo reconocimiento de entidades, normalización y desambiguación.

Otra tarea relevante es propuesta por May and Priyadarshi [22] en Semeval 2017, centrada en el reconocimiento y la generación de AMR a partir de oraciones médicas en inglés. La aplicación de una conceptualización de propósito general, como AMR, a dominios específicos alentó a los participantes a cerrar la brecha entre el desarrollo de técnicas generalizables y la aplicación de heurísticas específicas de dominio. Sin embargo, el reconocimiento del modelo AMR ya es un problema complejo en sí mismo, que puede tener un impacto negativo en la participación de los investigadores en estos eventos si no están especializados en este modelo. Los modelos más simples y de propósito general pueden alentar un mayor grado de participación dada una curva de entrada más fácil. Un ejemplo de esto último es el evento Semeval 2017 Task 10 [23], que propone la extracción de palabras claves y relaciones en documentos científicos, con un modelo simple basado en tres clases de entidades y dos relaciones de propósito general. Esta tarea recibió un número mucho mayor de presentaciones que la anterior, a pesar de que ambos eventos

se desarrollaron en el mismo contexto y se dirigieron a audiencias similares.

Fuera del marco de una competencia, los sistemas de evaluación abiertos y de larga duración permiten a los investigadores evaluar sus enfoques con métricas oficiales. Esto también puede proporcionar un repositorio centralizado del estado del arte, donde los enfoques existentes sean resumidos y enlazados a los artículos y recursos respectivos. En este sentido, esta investigación propone un sistema de evaluación en línea que permite una comparación de nuevos enfoques con resultados publicados oficialmente en cualquier momento. Sobre la base de esta infraestructura, se organizan en plazos programados campañas de evaluación oficiales con un diseño más competitivo.

1.3. Problema Científico

Los enfoques existentes para el descubrimiento automático de conocimiento en lenguaje natural tienen una aplicación limitada, debido a diversos factores. Por un lado, no existen suficientes recursos anotados, especialmente en idiomas diferentes del Inglés, necesarios para entrenar sistemas de aprendizaje automático. Además, los modelos de representación semántica existentes son específicos a un dominio o tarea concreta, mientras que los modelos generalizables son computacionalmente complejos de automatizar. Por otro lado, hay una fragmentación en la comunidad científica, con poca interacción entre comunidades que se concentran en enfoques específicos, tales como el aprendizaje profundo y la representación del conocimiento. Esta situación dificulta el desarrollo de técnicas capaces de descubrir conocimiento de propósito general en documentos de diferentes dominios e idiomas.

1.4. Objetivos

El objetivo general de esta Tesis es el diseño y construcción de un ecosistema para apoyar el desarrollo de tecnologías en el campo del descubrimiento de conocimiento. Este ecosistema consta de recursos lingüísticos, como la definición de un modelo semántico de anotación y corpus; herramientas e infraestructura para desplegar y evaluar sistemas; y, métricas de evaluación para permitir comparaciones justas. Concretamente, las contribuciones de

esta investigación son:

1. La definición de un modelo semántico y un esquema de anotación para capturar la semántica del lenguaje natural que es generalizable a cualquier dominio.

2. La construcción de varios recursos lingüísticos (corpus) manualmente anotados en idioma español, específicamente orientados al dominio de la salud, y un análisis de sus métricas de calidad y características fundamentales.

3. La definición formal de una tarea de descubrimiento de conocimiento basada en estos recursos lingüísticos, incluyendo métricas objetivas de evaluación en diferentes subtareas.

4. El desarrollo de una infraestructura para apoyar la creación de sistemas para la tarea mencionada, incluyendo herramientas y sistemas de base; y un servicio en línea para la evaluación automática y continua de nuevas técnicas.

5. La organización y evaluación de eventos competitivos para incentivar en la comunidad científica el desarrollo de tecnologías de descubrimiento de conocimiento en idioma español, así como un análisis de los resultados obtenidos y una discusión de las líneas de desarrollo más prometedoras.

6. El desarrollo de una estrategia computacional para acelerar la construcción de recursos lingüísticos mediante la inclusión de un sistema de aprendizaje en el ciclo de anotación que provea sugerencias al anotador humano.

1.5. Estructura de la Tesis

La presente tesis está organizada por el sistema de compendio de artículos científicos. Se presentan un total de seis artículos publicados que resumen todo el trabajo realizado en la consecución de los objetivos definidos en la Sección 1.4. La Parte I presenta en el Capítulo 2 un análsis del estado actual de la técnica en varios campos de investigación relevantes para esta Tesis. A modo de resumen se presenta además el Capítulo 3 que recoge los principales resultados obtenidos. Los artículos publicados organizan en la Parte II.

Finalmente se presentan unas conclusiones generales y recomendaciones de trabajo futuro en la Parte III.

A continuación se describe en mayor detalle el contenido de cada capítulo.

Parte I resume los resultados de la Tesis.

Capítulo 2 presenta un análisis detallado del estado de la técnica en tres campos de investigación relevantes para esta Tesis: la representación computacional del conocimiento, la anotación de recursos lingüisticos en lenguaje natural y la aplicación del aprendizaje automático en tareas de descubrimiento de conocimiento. Este capítulo profundiza en el estudio de los temas presentados en la Sección 1.2, así como otros conceptos, técnicas, y recursos relacionados.

Capítulo 3 presenta los principales resultados obtenidos en el curso de la investigación, agrupados en 6 secciones fundamentales dedicadas a cada uno de los objetivos de investigación presentados en la Sección 1.4.

Parte II presenta los seis artículos que conforman el compendio de investigación de esta Tesis, y que dan soporte a cada una de las contribuciones presentadas en la Sección 1.4:

Capítulo 4 da soporte a la Contribución 1, con el diseño conceptual de un esquema de anotación independiente de dominio e idioma para la extracción de conocimiento en lenguaje natural, presentado en el artículo *A General Purpose Annotation Model for Knowledge Discovery: A Case Study in Spanish Clinical Text*. Este esquema fue utilizado para la anotación de tres versiones del corpus *eHealth-KD* que será presentado en los capítulos siguientes.

Capítulo 5 da soporte a la Contribución 2, presentando la primera versión del corpus *eHealth-KD*, anotado en el contexto de la competencia *eHealth Knowledge Discovery* organizada en el Taller de Análisis Semántico (TASS 2018), descrita en el artículo *A Corpus to support eHealth Knowledge Discovery technologies*. Esta primera versión del corpus fue anotada con una versión simplificada del esquema presentado en el Capítulo 4.

Capítulo 6 da soporte a las Contribuciones 3 y 5, describiendo la primera edición del evento *eHealth Knowledge Discovery* a través del artículo

Analysis of eHealth Knowledge Discovery Systems inthe TASS 2018 Workshop. En este evento se utilizó la primera versión del corpus *eHealth-KD* presentado en el Capítulo 5. El artículo presenta además las métricas de evaluación, y una breve descripción de los sistemas que participaron en la competencia.

Capítulo 7 da soporte a la Contribución 4, que es el resultado principal de esta tesis, consistente en una segunda versión del corpus anotado con el esquema completo, así como el diseño de una infraestructura y recursos para el desarrollo de tecnologías de descubrimiento de conocimiento. Estos resultados se presentan en el artículo *A Computational Ecosystem to Support eHealth Knowledge Discovery Technologies in Spanish*.

Capítulo 8 da soporte a la Contribución 5, resumiendo la última edición hasta la fecha de este evento en el artículo *Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021*. Este artículo permite apreciar la evolución tanto de los recursos y tareas como de los enfoques presentados por los participantes a lo largo de cuatro ediciones de dicha competición.

Capítulo 9 da soporte a la Contribución 6, proponiendo una estrategia para acelerar considerablemente el proceso de anotación manual a partir de introducir un sistema de aprendizaje en el ciclo de anotación. Dicha estrategia se presenta en el artículo *Active Learning for Assisted Corpus Construction: A Case Study in Knowledge Discovery from Biomedical Text*. Esta estrategia fue desarrollada a partir de las experiencias ganadas durante el proceso de anotación, normalización y publicación del corpus *eHealth-KD* y los eventos competitivos relacionados.

Parte III presenta las conclusiones y recomendaciones.

Capítulo 10 presenta las conclusiones generales de la investigación, las experiencias adquiridas, y un resumen de las publicaciones que tributan a esta Tesis.

Capítulo 11 presenta las principales recomendaciones para dar continuidad a la investigación presentada en esta Tesis.

Capítulo 2

Estado del Arte

En las últimas décadas el crecimiento de Internet ha puesto a disposición de los seres humanos un volumen cada vez mayor de información. Este fenómeno ha sido denominado Big Data [24], y su estudio ha atraído la atención de diferentes comunidades de investigación como inteligencia empresarial [25], ingeniería que incluye física y biológica [26] y redes sociales [27]. Una fracción mayoritaria de esta información se encuentra en formato no estructurado: texto en lenguaje natural, imágenes, videos, audio, entre otros. A diferencia de la información estructurada (e.j., almacenada en bases de datos), la información no estructurada presenta ambigüedad, redundancia, y otras características que dificultan su comprensión. Esta dificultad es todavía mayor si se pretende construir sistemas computacionales que hagan uso del conocimiento implícitamente almacenado en grandes volúmenes información no estructurada. Por lo tanto, un primer paso en la utilización de estas fuentes de información consiste en representar la información existente en estructuras semánticas que sean computacionalmente tratables.

Las ontologías son una de las representaciones más comunes del conocimiento en formato digital [28]. El auge de la web semántica ha fomentado la creación de ontologías y bases de conocimiento a gran escala en varios dominios. Algunas ontologías representan dominios generales, como DBPedia, y recopilan una gran proporción del conocimiento humano común. Otros operan en un dominio más específico pero incorporan datos más detallados sobre el dominio. Uno de los mayores obstáculos para representar grandes

volumenes de conocimiento en estructuras semánticas es la cantidad de esfuerzo y tiempo que necesitan los expertos para construir ontologías de forma manual [29, 30]. En este sentido, el campo del aprendizaje de la ontología [31] estudia las técnicas y metodologías que permiten la extracción automática o semiautomática de ontologías de fuentes de información no estructuradas, como el texto en lenguaje natural [32, 33].

Descubrir automáticamente conocimiento relevante en lenguaje natural requiere del diseño de algoritmos que sean capaces de asociar una interpretación semántica a un fragmento de texto en lenguaje natural. La complejidad del lenguaje natural hace impracticable el definir de manera explícita las reglas que permiten a un ser humano comprender toda la variedad lingüistica existente. La alternativa es utilizar algoritmos que sean capaces de descubrir automáticamente estas reglas (o aproximaciones de las mismas) directamente de los datos. El aprendizaje automático es un área de la inteligencia artificial, de creciente auge, que permite diseñar algoritmos que aprendan a partir de datos [34]. Uno de los paradigmas más utilizados en este campo es el aprendizaje supervisado. En este paradigma, en vez de codificar directamente las reglas, se le provee a un algoritmo con un conjunto (generalmente grande) de ejemplos de entrada y salida, lo que permite realizar un proceso de entrenamiento en el cuál se infieren automáticamente las posibles reglas que mejor se ajustan a esos ejemplos [35]. Otros enfoques de aprendizaje automático, incluyendo no supervisado y por refuerzo, también pueden ser aplicados al descubrimiento de conocimiento, pero son menos comunes.

De manera general no es factible diseñar un algoritmo que obtenga directamente de lenguaje natural una ontología formalmente definida. Por el contrario, el proceso generalmente se divide en varias fases que permiten ir refinando progresivamente la información hasta extraer el conocimiento relevante. El primer paso consiste en detectar en el texto las entidades relevantes y las relaciones semánticas entre ellas. Este paso ya representa un reto significativo, pues la propia naturaleza del lenguaje natural hace que el mismo conocimiento pueda ser expresado de diferentes maneras, y que el mismo texto corresponda a diferentes interpretaciones. Resolver este problema con aprendizaje supervisado requiere de un número considerable de ejemplos de texto en lenguaje natural, donde un experto humano ha anotado los fragmentos que corresponden a entidades y relaciones.

Más allá del costo de este proceso de anotación manual, es necesario diseñar esquemas de anotación que sean a la vez suficientemente sencillos como para ser aprendidos automáticamente, y suficientemente expresivos como para ser capaces de representar la semántica deseada. Según el dominio de interés, es posible que las entidades y relaciones a detectar sean de tipos muy específicos (e.j., fármacos y enfermedades [3] o interacciones entre genes y proteínas [4]), o de propósito general.

Definir un esquema de anotación generalizable a muchos dominios e idiomas requiere encontrar una representación semántica común capaz de expresar ideas muy generales independientemente de la sintaxis y las reglas gramaticales de cada idioma. Una de las formas más comunes de representar piezas arbitrarias de conocimiento es a través de la noción de conceptos y acciones [36]. Al comprender el mundo como compuesto de conceptos que interactúan entre sí a través de acciones, los humanos pueden comunicar una gran variedad y complejidad de información y conocimiento. La forma más sencilla de combinar conceptos y acciones es a través de relaciones binarias, donde un concepto actúa como sujeto y otro como objetivo. Además, existen algunos tipos de relaciones suficientemente generales, como los hipónimos y los holónimos, que son utilizados en todos los dominios del discurso humano.

El proceso de descubrimiento de conocimiento a partir de lenguaje natural puede verse entonces como un flujo compuesto por varias etapas, que comienza con el texto y termina en una representación semántica del conocimiento relevante en forma de ontología. El primer paso consiste en la construcción manual o semi-automática de recursos lingüísticos anotados. Esto requiere elegir o definir un esquema de anotación que sea propicio para el dominio de interés. A partir de un corpus anotado, se procede al entrenamiento de algoritmos de aprendizaje automático que permitan aplicar el mismo esquema de anotación a grandes volúmenes de texto. Posteriormente, todas las entidades y relaciones descubiertas automáticamente son agrupadas en un grafo semántico. En este punto es posible realizar tareas de post-procesamiento para eliminar redundancias, combinar entidades semejantes, o detectar inconsistencias. Finalmente se obtiene una estructura semántica unificada, que puede ser en el formato de una ontología, donde se representa el conocimiento relevante que estaba implícito en el texto original.

Para entender el estado del arte en el descubrimiento automático de

conocimiento a partir de lenguaje natural es necesario analizar varias áreas de investigación. Las técnicas para la representación semántica de conocimiento, su almacenamiento y procesamiento computacional, y sus métricas de evaluación, se presentan en la Sección 2.1. Luego, la Sección 2.2 resume los principales esquemas de anotación existentes, así como las herramientas de anotación más populares. A continuación, la Sección 2.3 presenta una breve descripción y comparación de varios recursos lingüísticos y semánticos de múltiples dominios. Finalmente, en la sección 2.4 se analiza el papel que juega el aprendizaje automático en el descubrimiento de conocimiento, tanto desde la creación de modelos de detección de entidades y relaciones como desde la construcción de herramientas de anotación asistida más inteligentes.

2.1. Representación del Conocimiento

Desde los albores de la informática, uno de los problemas que ha atraído gran atención es el de representar el conocimiento en un formato computacional, de modo que se pueda realizar un razonamiento automático para descubrir verdades nuevas, previamente desconocidas [37]. Podría decirse que la tecnología de representación del conocimiento más popular en uso son las ontologías [28], que se han convertido en el estándar *de facto*. Las ontologías se pueden definir como una especificación formal de una conceptualización [38]. Esto representa conceptos, relaciones entre estos conceptos, instancias de estos conceptos y reglas de inferencia para derivar nuevas relaciones.

Como tales, las ontologías pueden considerarse como una combinación de dos enfoques predominantes para la representación del conocimiento: los basados en la lógica formal [39] y los basados en grafos de relaciones semánticas [40]. En los enfoques basados en la lógica, los hechos se representan como predicados o funciones lógicas y el razonamiento se habilita mediante la aplicación de reglas formales de inferencia. Por el contrario, las representaciones basadas en grafos expresan hechos como nodos (objetos) y aristas (relaciones) y el razonamiento se basa en los métodos de búsqueda en grafos. Sin embargo, en las ontologías, los objetos y sus atributos y relaciones se representan en un grafo de conceptos, que también se puede interpretar como un conjunto de predicados y funciones sobre estos objetos. Además de esta capa, se pueden agregar reglas de inferencia, que permiten utilizar métodos de razonamiento lógico para derivar nuevos atributos y relaciones

entre conceptos existentes.

Las relaciones en una ontología pueden ser de un dominio específico, pero a menudo se representan algunas relaciones de dominio generales, como *is-a* y *part-of*. Este tipo de relaciones permiten representar conceptos más abstractos o complejos a partir de la composición de conceptos más concretos o simples. Por tanto, muchas ontologías contienen algún tipo de taxonomía de conceptos cada vez más abstractos, que también están interconectados entre sí utilizando otras relaciones semánticas que pueden ser específicas de dominio. Estos recursos permiten representar marcos de conocimiento complejos, hasta un grado de especificidad que permite el diseño de herramientas de razonamiento totalmente automatizadas. Debido a la alta complejidad de los conceptos y relaciones que se representan, y la experiencia necesaria para reconocer los conceptos más relevantes de un dominio, las ontologías suelen ser construidas manualmente por expertos del dominio [41]. Así, construir una ontología es un proceso que requiere mucho tiempo y una gran cantidad de expertos para definirlo y poblarlo con instancias relevantes que se refieren a objetos y relaciones. Esto hace realmente difícil construir ontologías artesanales y asegurar su mantenibilidad, debido a que día a día aparece en la World Wide Web una gran cantidad de información nueva y valiosa, deseable para convertirla en conocimiento. Otro resultado importante de este proceso es que los expertos generalmente representan solo hechos que son absolutamente ciertos en el dominio. Aunque los formatos de ontología existentes pueden extenderse para tratar con [42] difusos o datos vagos [43], asignar manualmente un grado de creencia a un hecho específico es una tarea compleja.

Es posible distinguir entre dos tipos de ontologías: dominios generales (u ontologías superiores) y dominios específicos (o simplemente ontologías de dominios). Las ontologías específicas de dominio son aquellas que se ocupan de los conceptos y relaciones del dominio de conocimiento particular. Como ejemplos, podemos citar ontologías en las ciencias médicas [44, 45], o el campo de la ingeniería de software [46]. Otras ontologías son más generales, ya que pueden usarse en diferentes dominios, o se usan para tareas de propósito general que se emplean en muchas áreas. `WordNet` [47] es una ontología de propósito general que contiene la mayoría de las palabras del idioma inglés y las relaciones sintácticas y semánticas entre ellas. Se utiliza en muchas tareas de procesamiento de lenguaje natural y minería de texto. `DBPedia` [48] es

una ontología enciclopédica que contiene parte del conocimiento presente en Wikipedia[1]. Relaciona personas, eventos históricos, hechos, ubicaciones y otros conceptos, en un formato estructurado y consultable. Dado que las ontologías tienen una forma unificada de representar un solo hecho, concepto o relación utilizando el Localizador Uniforme de Recursos (URL), es posible y muy común que diferentes ontologías se vinculen entre sí. Por ejemplo, muchas ontologías específicas de dominio tienen entidades que están vinculadas a la entrada correspondiente en `DBPedia`. El enfoque de vincular y referenciar a otras ontologías ampliamente conocidas, conocido como *linked data* [49], permite estandarizar la representación del conocimiento compartido y facilita las tareas de consulta y análisis.

Las ontologías son una herramienta eficaz para representar el conocimiento en una amplia variedad de dominios y escenarios [50]. Son lo suficientemente flexibles para adaptarse a un dominio particular y lo suficientemente potentes como para representar conceptos complejos. Sin embargo, una de las tareas más complejas en este sentido es mantener una ontología actualizada con respecto a la masiva cantidad de datos no estructurados que se generan y publican todos los días. Por tanto, surge la necesidad de herramientas computacionales para construir ontologías con procesos automatizados o semiautomatizados.

2.1.1. Aprendizaje de Ontologías

El problema de descubrir, almacenar y utilizar el conocimiento en una forma computacionalmente eficaz ha sido ampliamente estudiado [32, 51, 52]. Este problema ha sido tratado desde dos áreas de investigación distintas pero complementarias: los campos de representación del conocimiento y aprendizaje automático. La comunidad de representación del conocimiento proporciona medios para representar y operar computacionalmente con el conocimiento almacenado en formas que pueden asegurar cierto grado de consistencia lógica. Por el contrario, la comunidad de aprendizaje automático proporciona herramientas para obtener conocimientos útiles a partir de grandes colecciones de datos estructurados y no estructurados. En la intersección del aprendizaje automático y la representación del conocimiento, ha surgido el campo del aprendizaje de ontologías para lidiar con la complejidad de

[1] http://www.wikipedia.org

mantener y actualizar ontologías manualmente. Este campo extrae técnicas y herramientas de ambas comunidades, para automatizar parte del proceso de creación y mantenimiento de ontologías.

El aprendizaje de ontologías tiene el potencial de reducir el costo de crear y, lo que es más importante, mantener ontologías grandes y complejas [52]. Este problema también se aborda en el paradigma de *learning by reading* [53], un campo en el que se utilizan técnicas del procesamiento del lenguaje natural y de la comunidad de representación del conocimiento. El propósito es construir una representación formal de algún campo en particular dados datos textuales sin restricciones relacionados con el campo. Esta representación también debe permitir un razonamiento completamente automático. El aprendizaje mediante la lectura se puede considerar como un caso particular de aprendizaje de ontología, aunque solo se refiere a la entrada textual, y la salida no está necesariamente en el formato de una ontología.

Se han propuesto varias herramientas y sistemas para esta tarea. `Text2Onto` [54] es un marco para el descubrimiento de cambios guíado por datos que emplea un modelo de ontología probabilística (POM). `OntoLT` [55] extrae conceptos y relaciones automáticamente de colecciones de texto anotadas lingüísticamente, por medio de un conjunto de reglas, que asigna clases lingüísticas a clases de ontología. Un enfoque más nuevo es `OntoGain` [56], que utiliza un enfoque no supervisado y explota la información léxica inherente del término de varias palabras para extraer conceptos de nivel superior.

En el campo del aprendizaje de ontologías, se pueden distinguir dos tareas generales de alto nivel: población de ontologías y enriquecimiento de ontologías [30]. La población de ontologías se ocupa del subproblema de encontrar nuevas instancias para una ontología ya definida, mientras que el enriquecimiento de ontologías se ocupa de agregar nuevos conceptos y relaciones a una ontología existente. Existe una superposición entre estas tareas y la mayoría de los enfoques existentes no pueden clasificarse únicamente en estos términos. En este campo, se han propuesto varias herramientas, que combinan diferentes enfoques y resuelven diferentes subconjuntos de las tareas de aprendizaje de ontologías. A continuación se hace una breve revisión de estos sistemas para ayudar a definir las principales características.

Los primeros enfoques, como `SYNDIKATE` [57], tratan solo de poblar una base de conocimiento, con una estructura ontológica predefinida (clases y

relaciones). Dado que la web es una rica fuente de información, varios enfoques se han centrado en extraer conocimiento de ella, explotando el formato semiestructurado de los recursos web. Algunos sistemas como `ARTEQUAKT` [58] y `SOBA` [59] son de dominio específico, centrándose respectivamente en el arte y los dominios deportivos. Otros sistemas, como `WEB->KB` [60] intentan construir bases de conocimiento de dominio general desde la web, explotando también la estructura de enlaces entre páginas para identificar relaciones. Otro ejemplo es el sistema `VIKEF` [61], que utiliza catálogos de productos como fuentes de datos, aprovechando así la estructura inherente presente en este tipo de datos. Aunque la mayoría de los sistemas intentan una extracción completamente automática, algunos ejemplos como `ADAPTATIVA` [62] incluyen una estrategia de inicialización, donde los expertos humanos brindan retroalimentación sobre el conocimiento extraído.

Para extraer el conocimiento relevante de un texto en lenguaje natural, se han introducido técnicas de PLN en sistemas como `OPTIMA` [63] e `ISODLE` [64]. El uso de características del lenguaje natural puede usarse para construir sistemas basados en reglas, como la propuesta `OntoLT` [55], que extrae conceptos y relaciones a través de un mapeo de clases lingüísticas a clases de ontología. Un enfoque alternativo es utilizar modelos estadísticos o probabilísticos, ejemplificados por sistemas como `LEILA` [65] o `Text2Onto` [54]. Otro ejemplo es `KnowItAll` [66], que introduce una métrica de información mutua puntual (PMI) para seleccionar instancias relevantes.

Una vez que las instancias de entidades y relaciones se extraen del texto, una pregunta natural es si se puede inferir un conocimiento más abstracto de estos ejemplos. Los sistemas que abordan este problema a menudo utilizan técnicas no supervisadas para intentar descubrir estructuras inherentes. Dos ejemplos relevantes de este enfoque son `OntoGain` [56] y `ASIUM` [67], que intentan construir automáticamente una jerarquía de conceptos usando técnicas de agrupamiento. El sistema `BOEMIE` [68] es otro ejemplo interesante, ya que intenta inferir automáticamente conceptos abstractos de las instancias concretas encontradas, pero se centra no solo en el texto, sino también en fuentes multimedia como imágenes y videos. La mayoría de los sistemas mencionados generalmente se enfocan en una iteración del proceso de extracción. Sin embargo, los enfoques más recientes, como `NELL` [32], intentan aprender continuamente de un flujo de datos web y aumentan con el tiempo tanto la cantidad como la calidad del conocimiento descubierto.

Uno de los problemas con muchos de estos enfoques es la cantidad de información falsa que generan [69]. En general, habrá muchas piezas de información redundantes o sin importancia en los corpus analizados. Un enfoque ingenuo que no tenga en cuenta este tema creará ontologías inmensas con muy poca información útil. Para abordar este problema, OntoGain propone un esquema de agrupamiento jerárquico que intenta identificar conceptos y relaciones generales.

En general, estas herramientas están enfocadas a la extracción de conocimiento y a la tarea de encontrar conocimiento relevante. Cuando se extrae conocimiento de una fuente confiable, incluso si es una fuente de lenguaje natural, tiene sentido enfocarse en optimizar el recobrado, es decir, obtener la mayor cantidad de información posible. Si la fuente de entrada es un conjunto de artículos médicos o la página web principal de una institución, existe una alta probabilidad de que la mayor parte de la información presente en esos documentos sea correcta. Por lo tanto, un procedimiento de extracción de ontología que maximice el recobrado obtendrá buenos resultados.

Sin embargo, cuando la fuente de entrada es de menor calidad, como blogs o publicaciones en redes sociales, existe una mayor probabilidad de que parte, o incluso la mayoría, de la información sea falsa o incorrecta. Si consideramos también el llamado fenómeno de la *posverdad*, y reconocemos que algunos autores comparten deliberadamente noticias o hechos falsos, el problema se vuelve mucho más difícil y urgente. Incluso si las mentiras deliberadas no fueran un problema, la mayor parte de la información compartida en las redes sociales y fuentes similares es irrelevante a largo plazo. En este contexto, el problema de extraer una ontología útil de un gran corpus de fuentes de Internet se vuelve menos un problema de reconocimiento de las piezas de información que se encuentran en el corpus, y más un problema de filtrado y selección de la información relevante, una vez extraída.

A pesar de la existencia de algunos sistemas de propósito general, no existe una propuesta definitiva que pueda aprender de manera simultánea y continua de las más variadas fuentes de información en línea. Otro desafío en este aspecto es obtener una representación computacionalmente conveniente de este conocimiento, independientemente del dominio, fuente y formato de los datos de entrada. Por último, un sistema de aprendizaje de ontología moderno tendrá que lidiar explícitamente con la gran cantidad de información

irrelevante o deliberadamente falsa que se difunde a través de fuentes en la web.

2.1.2. Evaluación de Sistemas de Descubrimiento de Conocimiento

Para evaluar un sistema de descubrimiento de conocimiento es necesario considerar dos dominios de características de interés. En primer lugar, como sistema computacional, existen métricas de ingeniería de software relevantes. Adicionalmente, siendo el principal objetivo de este tipo de sistemas la construcción de representaciones semánticas del conocimiento de un dominio concreto, es necesario evaluar también dicho conocimiento.

Desde el punto de vista de la ingeniería de software, es importante diseñar sistemas modulares y extensibles, de modo que puedan adaptarse fácilmente a los nuevos formatos de entrada, o nuevos algoritmos se pueden conectar e integrar fácilmente en todo el flujo. Además de estas métricas cualitativas, cada una de las tareas realizadas por un sistema se puede evaluar por separado de forma cuantitativa. La mayoría de estas tareas tienen una métrica de rendimiento definida que se puede utilizar para evaluar el grado de correctitud de dicha tarea. Para muchas de las tareas descritas en las secciones anteriores, existen métricas de rendimiento estándar en la literatura que pueden usarse para evaluar cada proceso en particular, tales como precisión, recobrado, F-medida, Kappa, entre otras.

Una métrica agregada de estos rendimientos individuales podría proporcionar una descripción general de alto nivel del desempeño de todo el sistema. Sin embargo, diseñar una métrica agregada que proporcione una medida práctica e interpretable de la calidad del rendimiento de todo un proceso de aprendizaje ontológico es una tarea compleja. Cada una de las diferentes tareas realizadas por el sistema puede tener un rendimiento de referencia muy diferente. Una precisión del 90 % puede ser un muy buen resultado en algunas tareas complejas, como el análisis de dependencias [70], pero mediocre en otras tareas, como la clasificación de imágenes [71]. Además, este número de referencia puede variar no solo entre las tareas, sino también en la misma tarea, según el conjunto de pruebas (o corpus) que se utilizce.

Subiendo un nivel de abstracción, para el problema general del aprendizaje

de la ontología, también hay varias métricas y metodologías de evaluación disponibles, como `OntoRand` [72] y `OntoMetric` [73]. Sin embargo, la mayoría de estas metodologías están diseñadas para evaluar una sola ontología que se crea o modifica utilizando técnicas de aprendizaje de ontologías. Extender estas metodologías a una colección de ontologías no es tan sencillo como agregar o promediar los resultados individuales. Por otro lado, cuando se trata de una colección de ontologías, pueden surgir otras inquietudes, como la consistencia intra-ontológica, que no suelen considerarse al evaluar una única ontología. A continuación se presentan algunos de los enfoques más comúnmente descritos en la literatura para evaluar ontologías [30].

M1- Comparación con un patrón oro. Este enfoque consiste en comparar una ontología aprendida con una ontología de referencia para el mismo dominio [74]. Se supone que la ontología de referencia es correcta y en gran medida representativa del dominio. Este método proporciona una gran compensación entre velocidad y precisión, ya que ambas ontologías se pueden comparar automáticamente en una serie de métricas sin intervención humana, y los resultados tienen una alta confiabilidad porque la ontología de referencia es creada por expertos. Existen algunas desventajas, por ejemplo, no siempre es fácil encontrar una buena ontología de referencia para un dominio dado, especialmente si el dominio no está muy bien definido o es muy novedoso. Por otro lado, incluso dos ontologías extraídas del mismo dominio por expertos pueden tener grandes diferencias con respecto a la estructura y, en particular, a los nombres que se asignan a las clases y relaciones. Esto requiere alguna forma de normalización y mapeo entre ambas ontologías antes de la comparación. Esta métrica es difícil de usar, especialmente cuando se crean nuevas características.

M2- Evaluación experta. Un término medio alternativo al enfoque anterior es tener un experto en el dominio (o varios) que analicen la ontología resultante y la evalúen de acuerdo con algunas métricas predefinidas [51]. Este es posiblemente el método más confiable, en el sentido de que proporciona el mayor grado de validación al que se podría aspirar. Sin embargo, la clara desventaja radica en la cantidad limitada de información que un ser humano puede procesar en un tiempo razonable. Esta desventaja se agrava en el caso en que se crea una ontología a partir de un corpus de datos muy grande,

como es el propósito de muchos sistemas que procesan texto en lenguaje natural. En este caso, se podría analizar un pequeño subconjunto de los datos y extrapolar los resultados, pero esta idea agrega la complejidad de determinar un subconjunto que es lo suficientemente relevante pero que sigue siendo de un tamaño manejable. Esta métrica suele ser costosa y difícil de usar.

M3- Evaluación a través de una aplicación. Un enfoque más práctico consiste en encontrar una aplicación interesante y evaluar si el uso de una ontología aprendida proporciona una mejora en esa aplicación [75]. Por ejemplo, usar una ontología aprendida sobre sentimientos humanos y frases relacionadas para mejorar el desempeño de un problema estándar de minería de opiniones. Si el uso del conocimiento representado en la ontología proporciona un incremento del rendimiento, medido por el enfoque estándar en la aplicación dada, esto proporciona una validación confiable de que el proceso para aprender la ontología, al menos, tiene un beneficio práctico medible. En cierto sentido, esta es una de las evaluaciones más valiosas para realizar, porque proporciona una línea base de comparación inmediata para un problema práctico. Los métodos anteriores que solo evalúan la ontología internamente no garantizan necesariamente que su contenido sea útil, incluso si es correcto según todas las métricas. Otra ventaja es que el proceso de evaluación se puede automatizar por completo y escalar para que coincida con la complejidad y el tamaño de la aplicación de destino. Como desventaja, validar un caso de uso no es necesariamente una métrica de la calidad general del conocimiento aprendido, y no está claro si esos resultados se replicarán en diferentes dominios y aplicaciones. Aunque esta métrica parece la más efectiva, en muchos casos al diseñar un sistema de descubrimiento de conocimiento, no se tiene definida una única aplicación de interés, sino un conjunto amplio de aplicaciones potenciales.

M4- Evaluación basada en datos. Finalmente, se puede realizar una evaluación basada en datos, comparando las entidades y relaciones en una ontología con un corpus de datos, no usados durante la construcción de la ontología, pero representativos del mismo dominio [76]. La ontología se puede evaluar contando el número de entidades superpuestas presentes en ella con las que se encuentran en el corpus. Se debe tener cuidado para permitir alguna

variación en el corpus con respecto a la ontología, por ejemplo, usando alguna forma de expansión de consultas. Este enfoque se ha utilizado para comparar relativamente diferentes ontologías creadas por expertos con el mismo corpus y decidir qué ontología proporciona el mejor ajuste con respecto al corpus [77]. Sin embargo, obtener una métrica absoluta de ajuste entre una ontología y un corpus es más difícil, principalmente porque no se sabe de antemano cuál es el valor de ajuste óptimo que se podría esperar. Otro posible problema de este enfoque, en el caso particular de las ontologías que se han aprendido de texto en lenguaje natural, es introducir inadvertidamente un sesgo en la evaluación. Si los métodos utilizados para comparar la ontología y el corpus de texto están correlacionados con los utilizados para construir la ontología, entonces los resultados serán de dudosa validez. Por ejemplo, si se usa un algoritmo NER durante la construcción de la ontología, y se usa el mismo algoritmo en el corpus para reconocer entidades relevantes; o si se usa alguna métrica de co-ocurrencia para detectar relaciones en ambos casos. Esta métrica es compleja de definir y en muchas ocasiones no es representativa.

Evaluar un solo método de aprendizaje de ontología es una tarea compleja, como lo demuestran los múltiples enfoques propuestos en la comunidad. Por lo tanto, es poco probable encontrar una única métrica automatizada para medir el rendimiento general de un sistema de descubrimiento de conocimiento de propósito general. El mejor enfoque parece ser utilizar una combinación de los métodos existentes, adaptados al escenario en cuestón. En algunos casos, se puede encontrar un patrón oro y utilizarlo para obtener una comparación de referencia. En otros casos, siempre que se agregue una interfaz adecuada para consultar fácilmente el conocimiento, un experto en el dominio puede interactuar con el marco y dar una evaluación cualitativa para el dominio de interés. Desde un punto de vista pragmático, la evaluación más interesante y valiosa parece ser encontrar problemas prácticos relevantes que se puedan resolver o mejorar al utilizar nuestro marco.

2.1.3. Comparación Cualitativa de Enfoques de Descubrimiento de Conocimiento

En la comunidad de aprendizaje de ontología, se han desarrollado varios marcos que atacan problemas de descubrimiento de conocimiento en varios

dominios. Algunos de los enfoques encontrados en la literatura se concentran en una tarea en particular, es decir, creación, población o enriquecimiento de ontologías, entre otras. Por ejemplo, marcos como `KnowItAll` [66], `Artequakt` [58] y `SOBA` [59] están orientados principalmente hacia la tarea de población de ontologías. Otros, como `ASIUM` [67], `VIKEF` [61] y `SYNDIKATE` [57] están orientados principalmente al enriquecimiento de ontologías. Sin embargo, muchas de las tareas o subproblemas que deben resolverse en cualquiera de estos dominios son muy similares y pueden reutilizarse. Por lo tanto, han surgido marcos más generales como `Text2Onto` [54] o `BOEMIE` [68] que tratan con una combinación de estas tareas.

En cuanto a la cantidad y complejidad de conceptos reconocidos, las soluciones existentes se pueden dividir en aquellas que solo extraen entidades (ej., `KnowItAll`), aquellas que solo extraen relaciones (ej., `ADAPTATIVA` [62], `LEILA` [65]) y aquellos que intentan extraer ambos (por ejemplo, `Artequakt`, `Web->KB` [60], `BOEMIE`). El descubrimiento de reglas de inferencia es otra tarea relevante que enriquece las ontologías ya construidas al agregar conocimientos de nivel superior, en forma de predicados lógicos o axiomas. Estos, a su vez, pueden usarse más adelante para descubrir instancias o relaciones faltantes, o para detectar valores atípicos y errores.

La mayoría de los sistemas emplean algún tipo de herramientas de aprendizaje automático para la mayoría de las tareas. En particular, muchos emplean herramientas de PLN para procesar texto natural y extraer conocimientos, y técnicas estadísticas para detectar agrupaciones. Sin embargo, en general, la arquitectura de estos sistemas generalmente sigue un flujo de trabajo estrictamente diseñado donde los componentes se conectan entre sí de formas predeterminadas.

La mayoría de los marcos y soluciones existentes requieren un cierto grado de interacción humana. En la mayoría de los casos, se espera que un experto en el dominio interactúe con una herramienta computacional para validar o refinar el resultado del proceso de aprendizaje. Sin embargo, esta posibilidad de interacción es un resultado del marco, no una necesidad.

Con respecto a las restricciones de dominio, podemos clasificar las soluciones existentes en aquellas que son completamente independientes del dominio (por ejemplo, `KnowItAll`, `LEILA`, `ISOLDE` [64]) y aquellas que están adaptadas a dominios particulares (por ejemplo, `SOBA`, `Artequakt`). Las so-

luciones independientes del dominio generalmente se diseñan de manera que no haya una dependencia particular ligada a un dominio, por lo tanto, son reutilizables en varios dominios. Sin embargo, esto significa que un sistema puede usarse en un dominio u otro, pero no significa necesariamente que el mismo sistema pueda aprender un poco de conocimiento de dos dominios diferentes *simultáneamente*.

Si un sistema está simplemente diseñado para ser independiente del dominio y se utiliza para aprender de dos dominios muy diferentes, el resultado esperado es una especie de ontología combinada que representa ambos dominios con una aproximación de la unión de los conceptos (entidades y relaciones) en ellos. Esta puede no ser la representación ideal, especialmente cuando se extiende a muchos dominios diferentes. Tratando de construir una ontología única que abarque todo el conocimiento que se puede extraer de varios dominios diferentes (posiblemente en el orden de cientos o miles) pueden ser significativamente más difíciles que una simple suma o unión de cada uno de los dominios individuales.

Es cierto que los humanos tenemos un conjunto básico de habilidades que podrían considerarse independientes del dominio, como nuestras habilidades innatas para la coincidencia de patrones. Estas habilidades básicas se utilizan en muchas de las tareas con las que nos enfrentamos a diario. La analogía computacional es un sistema con un único algoritmo de aprendizaje de propósito general que podría realizar, por ejemplo, tareas tan diferentes como el reconocimiento de voz, la clasificación de imágenes y la traducción con el mismo proceso. Este es el enfoque preferido por una parte de la comunidad de investigadores en aprendizaje automático [78] que esperan construir una inteligencia de propósito general a partir de un solo algoritmo de aprendizaje de propósito general y una única representación interna de propósito general.

Sin embargo, para tareas realmente complejas, se puede argumentar que los humanos usan representaciones especializadas, algoritmos de aprendizaje y técnicas de inferencia. Un experto humano en un dominio altamente complejo (como las matemáticas), no utiliza las mismas técnicas para la inferencia que las que se utilizan en tareas en tiempo real como el reconocimiento de objetos y de voz. La inferencia en dominios muy complejos no se puede realizar con herramientas basadas en la intuición. Por lo tanto, la construcción de un sistema inteligente que pueda realizar inferencias en varios dominios

altamente complejos simultáneamente requerirá el uso de representaciones y técnicas especializadas en cada dominio.

2.2. Anotación Semántica de Lenguaje Natural

Como ya se ha mencionado, el descubrimiento del conocimiento es un campo de la informática que muestra un crecimiento acelerado en muchos dominios, desde bases de datos [79, 80] hasta imágenes [81] y texto en lenguaje natural [82]. Específicamente en el texto en lenguaje natural, este campo es de gran relevancia en los dominios biomédico y de salud, donde se utiliza para realizar tareas como reconocimiento de entidades nombradas (NER), extracción de relaciones y generación de hipótesis, entre otros [83]. Estas tareas generalmente utilizan corpus anotados para aprender las características que aparecen en el texto y mapearlas con las estructuras de conocimiento. Para cada tarea, se han diseñado modelos de anotaciones específicos que se enfocan en elementos específicos del texto. Por ejemplo, en las tareas NER es más importante centrarse en frases nominales que en otras construcciones gramaticales.

Existen varios enfoques para construir representaciones semánticas del conocimiento para la anotación de texto en lenguaje natural. En muchos casos, estas representaciones utilizan una conceptualización específica de dominio. Aunque esto proporciona una representación más especializada, hace que estos enfoques sean más difíciles de aplicar a una amplia gama de dominios. Sin embargo, a pesar de que estas tareas específicas de dominio son diferentes, la mayoría de ellas comparten características comunes. Por ejemplo, la mayoría de las tareas se ocupan de la detección de entidades relevantes y sus relaciones. Por lo tanto, la creación de modelos de anotación de propósito general permite el diseño de técnicas de descubrimiento de conocimiento reutilizables y entre dominios. En esta línea, se han desarrollado varias representaciones semánticas independientes del dominio (por ejemplo, `AMR` [84], `PropBank` [16], `FrameNet` [85]). Estas representaciones se basan en gran medida en léxicos detallados que definen roles semánticos específicos para el significado de cada palabra. Por este motivo, desarrollar sistemas de descubrimiento de conocimiento con este nivel de detalle supone retos significativos.

Alternativamente, es posible utilizar una conceptualización de propósito general más simple, que sea capaz de representar entidades y hechos de múltiples dominios de conocimiento. Tal conceptualización debe ser lo suficientemente general como para dar cabida a dominios diferentes, siempre que se garantice el grado de expresividad necesario para las tareas de extracción de conocimiento. Una posible conceptualización de esta naturaleza es utilizar tripletas de *Sujeto-Acción-Objetivo* [86]. Esta estructura ha demostrado ser útil para representar el conocimiento en dominios específicos, como críticas de cine [86] o análisis de sentimiento. Además, las tripletas *Sujeto-Acción-Objetivo* extraídos automáticamente del texto se pueden vincular posteriormente a relaciones específicas de dominio mediante el uso de redes semánticas. Como ejemplo, el sistema `SemRep` [87] extrae tripletas *Sujeto-Predicado-Objeto* de textos en lenguaje natural en el dominio de la salud. Los predicados están vinculados a relaciones específicas en la red semántica `UMLS` [88].

El uso de representaciones semánticas más simples, incluso con la pérdida de cierta capacidad de representación, permite simplificar la creación de técnicas automáticas basadas en el aprendizaje automático. Estas representaciones también pueden usarse como la primera etapa en sistemas diseñados para una tarea específica de dominio, reutilizando así recursos y técnicas en dominios con pocos recursos disponibles.

Un trabajo reciente en el desarrollo de teleologías [36] sugiere que las tripletas *Sujeto-Acción-Objetivo* pueden ser la base para conceptualizaciones de propósito general en muchos dominios diferentes, ya que esta estructura permite la captura de interacciones entre objetos a través de las acciones que realizan. Un pequeño conjunto de relaciones semánticas, como *hiponomía* y *holonomía* pueden proporcionar una estructura semántica complementaria. Estas relaciones "generales" son comunes en la mayoría de las bases de conocimiento, independientemente del dominio, como `WordNet` [89], `DBPedia` [90] y `ConceptNet` [91]. Otras posibles conceptualizaciones permiten capturar la semántica del lenguaje natural, como la anteriormente mencionada *Representación de Significado Abstracto* (`AMR`) [84]. A pesar del poder de representación superior de `AMR` sobre estructuras simples como las tripletas *Sujeto-Acción-Objetivo* y relaciones semánticas básicas, el proceso de anotación para `AMR` es considerablemente más complejo tanto para humanos como para técnicas automatizadas.

La construcción de corpus anotados con la estructura *Acción-Sujeto-Objetivo* es el primer paso hacia el diseño de sistemas que puedan extraer automáticamente estas anotaciones. Existen varios corpus en la literatura, anotados con una variedad de esquemas diferentes, como `CLEF` [92], `YAGO` [14] y `EmotiNet` [33]. Sin embargo, la mayoría de estos recursos están anotados con conceptualizaciones específicas de dominio que son difíciles de extender a diferentes dominios de conocimiento.

2.2.1. Modelos de Anotación de Propósito General

Se han desarrollado varios modelos de anotación semántica de propósito general, que intentan representar la semántica de una oración más allá de la estructura sintáctica. Estos modelos se basan libremente en la estructura gramatical Sujeto-Verbo-Objeto que es omnipresente en el lenguaje humano. En esta sección se analizan cuatro modelos de representación semántica a nivel de oración basados en esta estructura gramatical.

FrameNet

`FrameNet` [85] es una base de datos léxica y un corpus anotado que modela los roles semánticos y las relaciones en una oración en lenguaje natural a través de estructuras conceptuales llamadas *frames*. Los marcos representan conceptos de propósito general, o eventos, que definen las posibles relaciones semánticas en las que esos conceptos pueden realizarse en lenguaje natural.

`FrameNet` se basa en una teoría del significado llamada *Frame Semantics*, derivada del trabajo de Fillmore et al. [93]. La idea básica es que los significados de la mayoría de las palabras pueden entenderse mejor sobre la base de un marco semántico, una descripción de un tipo de evento, relación o entidad y los participantes en él. Por ejemplo, el concepto de *cocinar* generalmente involucra a una persona que cocina (*Cocinero*), la comida que se va a cocinar (*Comida*), algo para sostener la comida mientras se cocina (*Recipiente*) y una fuente de calor (*Instrumento de calentamiento*). En el proyecto FrameNet, esto se representa como un marco llamado `Apply_heat`, y `Cook`, `Food`, `Heating_instrument` y `Container` se denominan elementos de marco (*frame elements*, FE). Las palabras que evocan este marco, como *freír*, *hornear*, *hervir* y *asar*, se denominan unidades léxicas (*lexical units*,

LU) del marco `Apply_heat`. Otros marcos son más complejos, involucrando más elementos, y otros son más simples, como *Colocar*, con solo un *Agente* (o *Causa*), una cosa que es colocada, y la ubicación en la que se coloca.

Muchos sustantivos comunes, como *árbol*, *sombrero* o *torre*, suelen servir como dependientes que encabezan un FE, en lugar de evocar claramente sus propios marcos. Por este motivo en `FrameNet` se dedica menos esfuerzo a anotar estos elementos, ya que la información sobre ellos está disponible en otros léxicos, como WordNet [89]. Sin embargo, dichos sustantivos también tienen una estructura de marco mínima propia y, de hecho, la base de datos `FrameNet` contiene un poco más de sustantivos que de verbos.

Formalmente, las anotaciones de `FrameNet` son conjuntos de tripletas que representan las realizaciones de FE para cada oración anotada, cada una de las cuales consta de un nombre de elemento de marco (por ejemplo, `Food`), una función gramatical (por ejemplo, *Objeto*) y un tipo de frase (por ejemplo, una frase nominal). Estos tres tipos de anotaciones en cada FE se pueden considerar como "capas" del esquema de anotación. La mayoría de las anotaciones son de oraciones separadas anotadas para una sola LU, pero también hay una colección de textos en los que se han anotado todas las palabras que evocan marcos; los marcos superpuestos proporcionan una rica representación de gran parte del significado de todo el texto. El equipo de `FrameNet` ha definido más de 1000 marcos semánticos y los ha vinculado mediante un sistema de relaciones de marcos, que relacionan marcos más generales con marcos más específicos y proporcionan una base para razonar sobre eventos y acciones intencionales.

Debido a que los marcos son fundamentalmente semánticos, a menudo son similares en todos los idiomas; por ejemplo, los marcos sobre compra y venta involucran al comprador, vendedor, bienes y dinero como elementos de marco, independientemente del idioma en el que se expresen. Se están llevando a cabo varios proyectos para construir FrameNets paralelos al proyecto `FrameNet` en inglés para otros idiomas, incluidos el español, alemán, chino y japonés, y se han llevado a cabo análisis y anotaciones semánticas de marcos en áreas especializadas, desde la terminología legal hasta el fútbol y el turismo.

VerbNet

`VerbNet` [94] es un léxico verbal que también define roles semánticos específicos para cada verbo. En `VerbNet`, los verbos se organizan en una jerarquía y se vinculan a través de diferentes roles temáticos, como agentes, causa, fuente o tema. Estos elementos permiten captar la representación semántica de oraciones.

`VerbNet` es un léxico jerárquico, independiente del dominio y de amplia cobertura de verbos con asignaciones a otros recursos léxicos, como WordNet, PropBank y FrameNet. El léxico stá organizado en clases de verbos que amplían las clases de Levin [95] mediante el refinamiento y la adición de subclases para lograr coherencia sintáctica y semántica entre los miembros de una clase. Cada clase de verbo está completamente descrita por roles temáticos, preferencias de selección de los argumentos y marcos que consisten en una descripción sintáctica y una representación semántica con estructura de subeventos modelada en el Modelo de Evento Dinámico propuesto por Pustejovsky [96].

Cada clase en `VerbNet` contiene un conjunto de descripciones sintácticas, o marcos sintácticos, que describen las posibles realizaciones superficiales de la estructura del argumento para construcciones como transitivo, intransitivo, frases preposicionales, resultantes y un gran conjunto de alternancias de diátesis. Las restricciones semánticas se utilizan para restringir los tipos de roles temáticos permitidos por los argumentos, y se pueden ,imponer restricciones adicionales para indicar la naturaleza sintáctica del constituyente que probablemente esté asociado con el rol temático. Los marcos sintácticos también pueden estar restringidos en términos de qué preposiciones están permitidas. Cada cuadro está asociado con información semántica explícita, expresada como una conjunción de predicados semánticos booleanos como "movimiento", "contacto" o "causa". Cada predicado semántico está asociado con una variable de evento E que permite a los predicados especificar cuándo en el evento el predicado es verdadero.

`VerbNet` se integra con clases de una extensión propuesta a la clasificación original de Levin [97]. Esta integración permite asociar descripciones sintáctico-semánticas detalladas a las clases de verbos nuevas, así como organizarlas apropiadamente en la taxonomía existente en `VerbNet`. El resultado es un recurso de libre acceso que constituye la clasificación de verbos más

completa y versátil para idioma inglés.

PropBank

`PropBank` [16] propone un esquema de anotación de propósito general, basado en predicados de anotación (verbos) como los principales constituyentes semánticos de una oración. El esquema de anotación de `PropBank` es capaz de representar varias relaciones semánticas, incluido el agente que causa una acción, el receptor de los efectos de una acción, modificadores de tiempo y ubicación y relaciones causales. Una característica clave de `PropBank` es que cada predicado define roles semánticos personalizados. Por ejemplo, el predicado " *accept* " define roles para el agente que acepta (`Arg0`), el objeto que se acepta (`Arg1`) y el agente de quien se acepta ese objeto. Los roles semánticos de `PropBank` son similares a los roles temáticos definidos en VerbNet y los elementos de marco en FrameNet. Como tal, existen recursos que vinculan estas estructuras semánticas [98]

El esquema de anotación de `PropBank` se desarrolló a partir de un consenso entre varios grupos de investigación. La primera fase de la anotación se centra en los predicados verbales, dejando a un lado los adjetivos, los sustantivos deverbales y los predicados nominativos para una etapa posterior. Con este propósito, se eligieron etiquetas de argumentos que pudieran mapearse fácilmente en las etiquetas utilizadas en la mayoría de las teorías modernas de estructura de argumentos sin estar especialmente relacionadas con ninguna teoría en particular. Por tanto, los argumentos de cada verbo se numeran como `Arg0`, `Arg1`, `Arg2`, etc., según la valencia del verbo en cuestión. El significado de cada etiqueta de argumento se define en relación con cada verbo en una colección de *frame files*. Cada conjunto de etiquetas de argumento y sus definiciones se denomina un *frameset* y proporciona un identificador único para el sentido del verbo, un significado para ese sentido del verbo y el conjunto de argumentos esperados. Por cada una de las definiciones, el esquema propone una serie de oraciones de ejemplo que demuestran varias realizaciones sintácticas para ese *frameset*.

Tomando como punto de partida el Corpus Penn Treebank II Wall Street Journal de un millón de palabras [99], `PropBank` agrega una anotación de estructura de argumento predicado. Al crear anotaciones para la estructura del argumento, se usa una combinación de factores sintácticos y semánticos,

aunque las claves sintácticas son las más importantes. El método general consiste en realizar un estudio de los usos de cada predicado dado, dividiéndose en sus sentidos principales si es necesario. Estos sentidos están divididos más por motivos sintácticos que semánticos, evitando así las divisiones detalladas de WordNet. Los argumentos esperados de cada sentido se numeran secuencialmente desde `Arg0` a `Arg5`. De acuerdo con pautas establecidas por la comunidad, no se intenta hacer que las etiquetas de los argumentos tengan el mismo "significado" de un sentido de un verbo a otro, por ejemplo, el papel que juega `Arg2` en un sentido puede corresponder al papel que juega `Arg3` puede en otro predicado dado, o en otro sentido.

AMR

Una propuesta más reciente es *Representación de Significado Abstracto* [84, AMR]. `AMR` constituye un esquema de representación semántica para oraciones en inglés que también intenta cubrir una amplia gama de relaciones semánticas con un modelo de propósito general. `AMR` incluye roles semánticos de `PropBank`, así como resolución de correferencia dentro de la misma oración, entidades y tipos nombrados, negación y otros modificadores en una estructura gráfica que representa el significado de una oración en lenguaje natural. Sin embargo, a pesar de que `AMR` captura el significado semántico completo de una oración, para el propósito del descubrimiento del conocimiento, sigue siendo considerablemente abstracto, y es necesario un procesamiento adicional para extraer estructuras concretas de conocimiento [100].

La estructura `AMR` consiste en un árbol etiquetado que debe ser fácil de leer para las personas y fácil de procesar computacionalmente. `AMR` tiene como objetivo abstraerse de las idiosincrasias sintácticas. Para ello, se intenta asignar el mismo árbol `AMR` a oraciones que tienen el mismo significado básico. Por ejemplo, a las frases "la describió como un genio", "su descripción de ella: genio" y "ella era un genio, según su descripción" se les asigna el mismo árbol `AMR`. `AMR` hace un uso extensivo de los conjuntos de marcos `PropBank`. Por ejemplo, una frase como "inversor en bonos" se representa utilizando el marco `invertir-01`, aunque no aparecen verbos en la frase. `AMR` es agnóstico al respecto de cómo derivar significados de cadenas, o viceversa. Al traducir oraciones a `AMR`, no existe una secuencia particular de aplicaciones de reglas ni alineaciones que reflejen dichas secuencias de reglas. Por otro lado, a pesar

de su generalidad, `AMR` está fuertemente sesgado hacia el inglés, y no es un recurso interlingua.

2.2.2. Herramientas de Anotación

Un elemento importante a considerar en la investigación sobre descubrimiento de conocimiento es la existencia de recursos e infraestructura computacionales que apoyan el desarrollo de nuevos enfoques. La creación de recursos lingüísticos a menudo surge de un proceso de anotación manual por parte de expertos humanos, que requiere herramientas computacionales para la anotación real, así como mecanismos para mezclar anotaciones y cálculo de métricas de consenso, idealmente en un entorno colaborativo. Una vez que se crean los recursos, es necesario distribuir el corpus, los modelos de referencia y las herramientas correspondientes entre la comunidad de investigación, a menudo a través de plataformas de intercambio de código fuente en línea.

En Neves and Ŝeva [1] se proporciona un análisis extenso y una comparación de varias herramientas de anotación. La tabla 2.1 resume las principales características que consideramos relevantes para esta investigación e identifica las herramientas de anotación más apropiadas entre un subconjunto de alternativas populares. Consideramos como requisitos indispensables herramientas de anotación de código abierto basadas en tecnologías web que permiten anotaciones de etiquetas múltiples, así como anotaciones de relación. El soporte para la anotación colaborativa, al menos parcialmente, también es muy deseable. De las herramientas analizadas, identificamos `Brat` [101] y `WebAnno` [102], ya que cumplen con todos los requisitos antes mencionados. En nuestra investigación, preferimos a `Brat` sobre `WebAnno` porque, aunque `WebAnno` ofrece más funciones, `Brat` permite una configuración más sencilla. No solo es más rápido iniciar un proyecto de anotación con esta herramienta, sino también capacitar a los anotadores para que utilicen su interfaz.

A continuación se resumen las características consideradas en esta comparación:

anotaciones multi-etiqueta: si la herramienta de anotación permite asignar más de una etiqueta a la misma secuencia de tokens, así como intersecciones entre diferentes secuencias anotadas con etiquetas diferentes. Esta característica es imprescindible si una misma frase o

Características	GATE Teamware	Knowtator	WebAnno	Brat	BioQRator	CATMA	Prodigy	TextAE	LightTag	Djangology	MyMiner	WAT-SL
anotaciones multi-etiqueta			✓	✓		✓			✓	✓		
anotaciones de relaciones		✓	✓	✓	✓			✓	✓		≈	
esquema personalizado	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
interfaz colaborativa	✓		≈	≈	≈	≈	≈		✓	✓		≈
interfaz web	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
auto-manejado	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓
código abierto	✓	✓	✓	✓		✓		✓		✓	✓	✓
referencia	[103]	[104]	[102]	[101]	[105]	[106]	[107]	[108]	[109]	[110]	[111]	[112]

Tabla 2.1: Comparación cualitativa de herramientas de anotación populares mencionadas en la literatura. Adaptada de la Tabla 3 in Neves and Ŝeva [1]. El símbolo ≈ indica que la característica correspondiente está parcialmente soportada.

palabra puede jugar más de un rol semántico en el esquema de anotación propuesto. En muchas herramientas de anotación se asume que cada token o frase tiene a lo sumo un rol semántico, por ejemplo, una entidad nombrada.

anotaciones de relaciones: si la herramienta permite anotar relaciones entre frases etiquetadas. Generalmente la anotación de una relacioón no corresponde a un elemento sintáctico de la oración, y debe ser anotada fuera de la superficie del texto. Por ejemplo, si se está anotando un evento y un agente que lo causa, es posible que la relación entre ambos esté implícita en la estructura sintáctica de la oración. Por ejemplo, en la oración "el asma inflama las vías respiratorias", el agente "asma" es causante del evento "inflama", pero la relación no puede ser atribuída a ninguna frase o palabra en particular, sino que se infiere a partir de la conjugación del verbo "inflamar". En estos casos, una herramienta que solo permita anotar elementos semánticos que correspondan a frases en la superficie del texto no será capaz de capturar esta relación.

esquema personalizado: si la herramienta permite configurar un esquema de anotación arbitrario. Varias herramientas existentes se diseñan para un esquema concreto, o permiten modificaciones menores (e.j., adicionar tipos de entidades nombradas). Las herramientas más flexibles permiten

definir completamente la estructura de las entidades y relaciones (en caso de soportar este tipo de anotación), así como las reglas que determinan qué entidades pueden asociarse mediante qué relaciones.

interfaz colaborativa: si la herramienta soporta la anotación simultánea de forma colaborativa entre más de un anotador. Generalmente este soporte requiere del uso de tecnologías web. Algunos de los sistemas que soportan algún tipo de colaboración existen restricciones sobre la cantidad de colaboradores que pueden estar en línea.

interfaz web: si la herramienta de anotación es accesible a través de una interfaz web. Esta característica generalmente va asociada a una mayor facilidad de configuración y uso, sobre todo si los anotadores no tienen la experticia técnica para configurar el proyecto de anotación. En muchos entornos, el especialista que configura el proyecto de anotación y el que anota tienen formación y habilidades diferentes, por ejemplo, informáticos y lingüistas respectivamente. Si la herramienta de anotación debe ser instalada en cada terminal donde será usada, se complejiza el proceso de introducir nuevos anotadores. Por otro lado, si la herramienta puede ser desplegada en un sistema computacional centralizado y accedido vía web por los anotadores, estos no tienen que lidiar con ninguna fase del proceso de instalación y configuración. Así mismo, esto significa que los datos se mantienen bajo un control centralizado, y facilita cualquier tarea de mezcla, copia de seguridad, y mantenimiento en general.

auto-manejado: si la herramienta de anotación puede ser instalada y manejada desde una infraestructura local, en vez de ser, por ejemplo, un servicio web provisto por un tercero. Las herramientas de escritorio por definición son auto-manejadas, pero varias herramientas web comerciales solo están disponibles desde una plataforma tercerizada. Esto significa que los datos a anotar deben ser subidos a la plataforma, lo que puede conllevar un riesgo de privacidad en algunos escenarios. Por otra parte, un servicio web provisto por un tercero será más fácil de utilizar, ya que todo el mantenimiento de la infraestructura es responsabilidad del proveedor del servicio. Algunas herramientas comerciales, como `Prodigy`, aunque no son de código abierto, aún así son auto-manejadas.

código abierto: si la herramienta tiene una licencia de código abierto. Esto

puede ser una característica deseada debido a consideraciones ya sea comerciales o de índole ética o científica. Las herramientas de código abierto, aunque pueden no contar con el nivel de soporte técnico que brindaría un proveedor de un servicio comercial, a menudo tienen una mejor documentación y una comunidad de usuarios mayor, lo que facilita su uso. Por otro lado, en principio estas herramientas pueden ser modificadas e incorporadas dentro de otros sistemas siempre que se respeten las restricciones legales que implica cada licencia.

A continuación se describen en mayor detalle las herramientas analizadas en esta investigación.

GATE Teamware

`GATE Teamware` es una herramienta de anotación de texto de código abierto y una metodología para la implementación y el apoyo de proyectos de anotación complejos. Tiene una arquitectura basada en tecnologías web, donde una serie de servicios web (por ejemplo, almacenamiento de documentos, anotación automática) están disponibles a través de HTTPS y los usuarios interactúan con las interfaces de anotación de texto a través de un navegador web estándar.

`GATE Teamware` se basa en `GATE` [113], una plataforma de PLN de código abierto extemsible, robusta y ampliamente utilizada. GATE contiene con numerosos componentes de procesamiento de texto reutilizables para muchos lenguajes, junto con un entorno de desarrollo gráfico de PLN e interfaces de usuario para la visualización y edición de anotaciones lingüísticas, árboles de análisis, cadenas de co-referencia y ontologías. Sin embargo, `GATE Teamware` fue creado específicamente para ser utilizado por anotadores no expertos, así como para permitir proyectos de anotación de corpus metodológicamente sólidos, eficientes y rentables en la web.

Además de sus usos de investigación, `GATE Teamware` también se ha probado como un marco para servicios de anotación comercial rentables, suministrados como unidades internas o como actividades especializadas subcontratadas. Se han llevado a cabo varios proyectos de anotación de prueba en los dominios de la bioinformática y la inteligencia empresarial, con una formación mínima y produciendo corpus de alta calidad. Sus resultados

muestran que el uso de `GATE Teamware` para la preanotación automática y la corrección manual aumenta la velocidad con la que se pueden procesar los documentos para su inclusión en la base de datos en un factor de alrededor del 50 %.

Al igual que otros softwares que requieren un servidor, la instalación de `GATE Teamware` es una tarea especializada, no trivial con costos asociados, en términos de tiempo significativo y experiencia del personal requerido. Para reducir esta barrera y proporcionar cero costos de inicio, los autores facilitan la elección de un conjunto de documentos anotados automáticamente y enviarlos a una instancia de `GATE Teamware`. También hay una distribución de máquina virtual que se puede descargar y ejecutar localmente.

Knowtator

`Knowtator` es una herramienta de anotación basada en Protégé [114]. En `Knowtator`, un esquema de anotación se define con definiciones de clases, instancias, ranuras y facetas de Protégé utilizando la función de edición de la base de conocimientos de este sistema. El esquema de anotación definido se puede aplicar a una tarea de anotación de texto sin tener que escribir ningún software específico para esa tarea o editar archivos de configuración especializados. Los esquemas de anotación en `Knowtator` pueden modelar fenómenos sintácticos (por ejemplo, análisis sintácticos superficiales) y semánticos (por ejemplo, interacciones proteína-proteína).

`Knowtator` aborda la definición de un esquema de anotación como una tarea de ingeniería del conocimiento aprovechando las fortalezas de Protégé como editor de una base de conocimiento. Protégé tiene componentes de interfaz de usuario para definir marcos de clases, instancias, ranuras y facetas. Un esquema de anotación de `Knowtator` se crea definiendo marcos utilizando estos componentes de interfaz de usuario como lo haría un ingeniero de conocimiento al crear un modelo conceptual de algún dominio. Para `Knowtator`, las definiciones de marcos modelan los fenómenos que la tarea de anotación busca capturar.

Una fortaleza clave de `Knowtator` es su capacidad para relacionar anotaciones entre sí a través de las definiciones de ranuras de las clases anotadas correspondientes. Las restricciones de las ranuras garantizan que las relaciones entre las anotaciones sean coherentes. Protégé es capaz de representar

modelos conceptuales mucho más sofisticados y complejos que otros sistemas, y que pueden ser utilizados por `Knowtator` para la anotación de texto. Además, debido a que Protégé se usa a menudo para crear modelos conceptuales de dominios relacionados con disciplinas biomédicas, `Knowtator` es especialmente adecuado para capturar entidades nombradas y relaciones entre entidades nombradas para esos dominios.

WebAnno

`WebAnno` es una herramienta de anotación basada en tecnologías web `WebAnno` que presenta varias funcionalidades que permiten la anotación de estructuras semánticas. En primer lugar, las características de ranura permiten el modelado apropiado de estructuras de predicado-argumento para etiquetado de roles semánticos (*Semantic Role Labelling, SRL*). `WebAnno` soporta los siguientes tipos de anotaciones semánticas adicionales: participantes y circunstancias para la anotación de eventos, relaciones n-arias para la extracción de relaciones y tareas de relleno de espacios para la extracción de información.

Además las restricciones ayudan a los anotadores al realizar un filtrado sensible al contexto de los conjuntos de etiquetas semánticas disponibles. Por ejemplo, el sentido de un predicado semántico determina los roles de argumento disponibles, como en el caso de esquemas tipo `PropBank`. Este filtrado es necesario para evitar perder un tiempo valioso al hacer que los anotadores busquen en una gran cantidad de etiquetas o que ingresen etiquetas manualmente. Las reglas de restricción se pueden definir manualmente o se pueden generar automáticamente, por ejemplo, a partir de recursos léxicos debidamente codificados.

Por último, `WebAnno` ofrece una interfaz de anotación mejorada para un proceso de anotación optimizado utilizando una barra lateral permanentemente visible en lugar de un cuadro de diálogo emergente para editar anotaciones y sus características. Estas nuevas funcionalidades se integran bien con las funcionalidades existentes en versiones anteriores de `WebAnno`, en particular su soporte para la anotación de estructuras sintácticas, permitiendo así la anotación semántica en coordinación con la anotación sintáctica.

`WebAnno` se desarrolló e implementó en coordinación con anotadores expertos en el contexto de un proyecto de anotación [115] para la desambiguación

del sentido de las palabras (WSD) y etiquetado de roles semánticos (SRL) en textos en idioma alemán. Algunos ejemplos de esquemas de SRL bien conocidos motivados por diferentes teorías lingüísticas fueron presentados anteriormente en este capítulo: `FrameNet`, `PropBank` y `VerbNet`. La anotación de SRL se basa típicamente en estructuras sintácticas obtenidas de bancos de árboles, como el Penn Treebank basado en constituyentes (para la anotación de `PropBank`), o el banco de árboles TIGER en alemán para la anotación de estilo `FrameNet` [116]. Un argumento se identifica típicamente por la extensión de su cabeza sintáctica o constituyente sintáctico. Para algunos esquemas de anotaciones (por ejemplo, `FrameNet`), la tarea también incluye desambigüación de sentidos.

Brat

`BRAT` es una herramienta de anotación basada en STAV, un visualizador de anotaciones de texto de código abierto Stenetorp2011b. `BRAT` fue fue diseñado para ayudar a los usuarios a comprender las anotaciones complejas que involucran una gran cantidad de tipos semánticos diferentes, anotaciones de texto densas, parcialmente superpuestas y conjuntos no proyectivos de conexiones entre anotaciones. Ambas herramientas comparten un componente de visualización basado en gráficos vectoriales, que proporcionan detalles y renderización escalables. `BRAT` integra la funcionalidad de exportación de formato de imagen PDF y EPS para admitir el uso en documentos científicos y publicaciones en formatos web.

`BRAT` amplia las capacidades de STAV implementando soporte para la edición de anotaciones, agregando funcionalidades para reconocer gestos de interfaz de usuario estándar que son familiares de los editores de texto, software de presentación y muchas otras herramientas. En `BRAT`, un tramo de texto se marca para anotación seleccionándolo con el ratón, arrastrando o haciendo doble clic en una palabra. De manera similar, las anotaciones se vinculan haciendo clic con el mouse en una anotación y arrastrando una conexión a la otra. Estos mecanismos de interacción hacen de `BRAT` una herramienta efectiva para anotadores con menor nivel de experticia técnica en el manejo de aplicaciones informáticas.

`BRAT` se basa en un navegador y se construye en su totalidad utilizando tecnologías web estándar. Por lo tanto, ofrece un entorno familiar para los

anotadores, ya que puede ser utilizado sin necesidad de instalar o distribuir ningún software de anotación adicional ni utilizar complementos del navegador. El uso de estándares web también hace posible que `BRAT` identifique de forma única cualquier anotación mediante los identificadores uniformes de recursos (URI). Esta funcionalidad permite permite vincular anotaciones individuales para discusiones en correo electrónico, documentos y páginas web, lo que facilita la comunicación con respecto a las anotaciones.

`BRAT` es completamente configurable y se puede adaptar a la mayoría de las tareas de anotación de texto. La primitiva de anotación más básica identifica un intervalo de texto y le asigna un tipo (o etiqueta), marcando tokens, fragmentos, o menciones de entidades nombradas. Estas anotaciones base se pueden conectar mediante relaciones binarias, dirigidas o no dirigidas, que se pueden configurar para, por ejemplo, extracción de relación simple, o anotación de marco de verbos al estilo `VerbNet`. `BRAT` admite también asociaciones n-arias de anotaciones, lo que permite la anotación de estructuras de eventos. Los aspectos adicionales de las anotaciones se pueden marcar utilizando atributos binarios o de valores múltiples que se pueden agregar a otras anotaciones. Finalmente, los anotadores pueden adjuntar notas de texto de forma libre a cualquier anotación.

Además de las tareas de extracción de información, estas primitivas de anotación permiten configurar `BRAT` para su uso en varias otras tareas, como fragmentación, etiquetado de roles semánticos y anotación de dependencias. Además, tanto el cliente `BRAT` como el servidor implementan soporte completo para el estándar Unicode. `BRAT` se distribuye con ejemplos de más de 20 corpus para una variedad de tareas, que involucran textos en siete idiomas diferentes e incluyen ejemplos de corpus como los presentados para las tareas compartidas de CoNLL sobre el reconocimiento de entidades con nombre independientes del idioma [117] y análisis de dependencia multilingüe buchholz2006conll. `BRAT` también implementa un sistema configurable para verificar restricciones detalladas en la semántica de la anotación, por ejemplo, especificando que un evento debe tomar obligatoriamente ciertos tipos de argumentos, y opcionalmente otros. La verificación de restricciones está integrada en la interfaz de anotaciones y la retroalimentación es inmediata, con efectos visuales que marcan las anotaciones incompletas o erróneas.

BRAT admite dos enfoques estándar para integrar los resultados de las herramientas de anotación automáticas en un flujo de trabajo de anotación. Las importaciones de anotaciones masivas se pueden realizar mediante herramientas de conversión de formato distribuidas con BRAT para varios formatos estándar (como BIO en línea y con formato de columna). Además, BRAT se puede integrar con herramientas que proporcionan interfaces de servicios web estándar configurables para ser invocadas desde la interfaz de usuario. Se debe tener en cuenta, sin embargo, que los juicios humanos no se pueden reemplazar o basar en un análisis completamente automático sin algún riesgo de introducir sesgos y reducir la calidad de las anotaciones. Para abordar este problema, BRAT propone mecanismos para aumentar el proceso de anotación con información de métodos estadísticos y de aprendizaje automático para respaldar el proceso de anotación y, al mismo tiempo, implicar el juicio del anotador humano para cada anotación.

BRAT implementa un conjunto de funciones de búsqueda que permite a los usuarios realizar búsquedas en columnas de documentos. Además, los resultados de la búsqueda pueden mostrarse opcionalmente usando la concordancia de palabras clave en el contexto y ser ordenados para navegar utilizando cualquier aspecto de la anotación coincidente (por ejemplo, tipo, texto o contexto). Una desventaja de BRAT radica en la imposibilidad de relacionar entre sí o asignar atributos a anotaciones de relaciones.

BioQRator

BioQRator es una interfaz de usuario de propósito general para anotar bioentidades y relaciones. Esto le permite a uno crear fácilmente una interfaz personalizada para cualquier proyecto de bio-curación si la tarea involucrada es anotar entidades y/o relaciones. Un tema importante para los sistemas de annotación son los múltiples formatos diferentes que se utilizan. Para abordar este problema, los autores adoptan BioC como formato estándar de entrada y salida. Para la entrada, se pueden utilizar documentos con formato BioC o resúmenes de PubMed. Para la salida, los documentos anotados también se pueden guardar en formato BioC o CSV (valores separados por comas). Más importante aún, BioQRator proporciona una interfaz web interactiva fácil de usar. También es compatible con varios navegadores, incluidos Chrome, Firefox y Safari (parcialmente compatible con Internet Explorer).

CATMA

CATMA es una herramienta desarrollada en la Universidad de Hamburgo que es utilizada actualmente por más de 60 proyectos de investigación en todo el mundo. CATMA ofrece una combinación de tres características principales que no se encuentran comúnmente en otras herramientas de análisis de texto:

1. Admite la anotación y el análisis colaborativos: un texto o corpus de texto puede ser investigado individualmente, pero también de forma conjunta por un grupo de estudiantes o investigadores.

2. Apoya prácticas exploratorias no deterministas de anotación de texto: un enfoque discursivo y orientado al debate para la anotación de texto basado en las prácticas de investigación de las disciplinas hermenéuticas es el modelo conceptual subyacente.

3. Integra la anotación de texto y el análisis de texto en un entorno de trabajo basado en la web, lo que hace posible combinar la identificación de fenómenos textuales con su investigación de forma iterativa y sin fisuras.

Lo que distingue a CATMA de otros métodos de anotación digital es su enfoque 'no dogmático': el sistema no prescribe esquemas o reglas de anotación definidos, ni obliga al usuario a aplicar taxonomías rígidas de sí o no, correctas o incorrectas a los textos (aunque también permite esquemas más prescriptivos). Más bien, la lógica de CATMA invita a los usuarios a explorar la riqueza y las múltiples facetas de los fenómenos textuales de acuerdo con sus necesidades: los usuarios pueden crear, expandir y modificar continuamente sus propios conjuntos de etiquetas individuales, por lo que si un pasaje de texto invita a más de una interpretación, nada en el sistema evita la asignación de anotaciones múltiples o incluso contradictorias. A pesar de toda esta flexibilidad, CATMA no produce anotaciones idiosincrásicas: todos los datos de marcado se pueden exportar en formato TEI / XML y reutilizar en otros contextos.

Dado que CATMA es una herramienta intuitiva, también es adecuada para humanistas con poco conocimiento técnico: la GUI permite un inicio rápido y el generador de consultas de CATMA (un widget basado en diálogos paso a paso) ayuda a los usuarios a recuperar información compleja de textos sin

tener que aprender un idioma de consulta. Otra ventaja en el lado fácil de usar es el hecho de que las funciones automatizadas de lectura a distancia de `CATMA` se mejoran y amplían continuamente: la versión actual ya presenta una serie de rutinas de anotación automatizadas, entre otras, la identificación de características narrativas básicas en los textos.

Prodigy

`Prodigy` es una herramienta de anotación comercial para crear datos de entrenamiento y evaluación para modelos de aprendizaje automático. También se puede utilizar `Prodigy` para ayudar a inspeccionar y limpiar los datos, realizar análisis de errores y desarrollar sistemas basados en reglas para usar en combinación con modelos estadísticos. Además de la interfaz visual contiene una biblioteca de código en Python.

La biblioteca incluye una variedad de flujos de trabajo prediseñados y instrucciones de línea de comandos para varias tareas, así como componentes para implementar los scripts propios del usuario en su flujo de trabajo. Los scripts pueden especificar cómo se cargan y guardan los datos, cambiar qué preguntas se hacen en la interfaz de anotación y definir HTML y JavaScript personalizados para cambiar el comportamiento de la interfaz. La aplicación web está optimizada para una anotación rápida, intuitiva y eficiente.

TextAE

`TextAE` es una herramienta de código abierto para la visualización y edición de anotaciones del proyecto PubAnnotation [118]. Como proyecto de código abierto, está disponible bajo la licencia de software MIT. `TextAE` provee una interfaz visual basada en tecnologías web y estándares abiertos, y soporta cualquier idioma representable en formato UTF8.

`TextAE` puede ser instalado en infraestructura propia, o utilizado como un servicio en línea sin requerir instalación. Esta funcionalidad facilita la interacción con anotadores con menor experticia en las tareas técnicas de instalación y configuración de la herramienta. Además, `TextAE` proporciona un componente REST para obtener anotaciones en formato `PubAnnotation JSON` desde la web, así como servir los documentos propios en este mismo formato.

`TextAE` soporta anotación de entidades nombradas y relaciones, con un esquema personalizable. Soporta además la anotación de atributos en las entidades. El motor de visualización de `TextAE` puede ser embebido en aplicaciones web mediante mecanismos de programación estándar, sin necesidad de componentes adicionales ni extensiones del navegador. Esto permite crear documentación, guías de anotación, y otros archivos en formato HTML con la misma visualización que los anotadores encontrarán en la herramienta.

LightTag

`LightTag` es una herramienta de anotación comercial que funciona a través de una interfaz web tanto manejada por el proveedor de la herramienta como auto-manejada. En este sentido, se diferencia de otras herramientas similares de código cerrado en que puede ser integrada con sistemas propios e instalada en infraestructura propia, en caso de ser necesario.

`LightTag` propone varias funcionalidades interesantes para aumentar la productividad de los anotadores. En primer ligar, permite anotaciones multi-etiqueta de entidades y relaciones, así como etiquetas a nivel de documento, tanto clases binarias como mútiples. Así mismo, es posible anotar frases, palabras exactas, o cualquier división a nivel de símbolos. A partir de una interfaz de búsqueda, permite la anotación de entidades usando listas de etiquetas largas. Esta herramienta permite crear guías de anotación dentro de la interfaz de anotación, y asociarla a los proyectos y documentos, de forma que sean fácilmente accesibles por los anotadores. Mediante una interfaz web, es posible asociar predicciones de modelos de aprendizaje a documentos en forma de pre-anotaciones, que luego pueden ser revisadas y corregidas por los anotadores.

Desde el punto de vista del manejo de proyectos, las versiones más avanzadas de `LightTag` permiten la anotación colaborativa, así como la creación de equipos y el manejo de las tareas de anotación con un sistema de calendario y asignación automática. En un proyecto de anotación, es posible definir roles para los diferentes anotadores, y crear reportes de rendimiento para comparar la efectividad de cada anotador o equipo. Finalmente, `LightTag` ofrece funcionalidades para el análisis y mantenimiento de la calidad de las anotaciones. Produce automáticamente métricas de consenso entre anotado-

res en tiempo real, y permite la corrección y verificación de anotaciones *a posteriori* en función de los roles definidos.

Djangology

La aplicación web de anotación `Djangology` se creó originalmente para satisfacer las necesidades de un proyecto de anotación colaborativo que involucra a más de 250 participantes internacionales. El objetivo del proyecto era crear un corpus estándar que esté anotado con entidades nombradas del dominio de interés: estudios médicos de las condiciones de trauma, shock y sepsis. Los resúmenes de una conferencia de la North American Shock Society se utilizaron para identificar las entidades nombradas de dominio específico a través de un proceso automatizado. Las anotaciones de entidades nombradas tuvieron que ser validadas por expertos en el dominio. El sistema `Djangology` estuvo en uso durante dos años consecutivos (2008 y 2009), logrando una tasa de respuesta media de los contribuyentes del 70 %.

Las necesidades del proyecto llevaron a un conjunto de requisitos comunes a proyectos similares de anotación colaborativa altamente distribuida. Se necesitaba una interfaz de administración para gestionar documentos y usuarios, así como para la definición de esquemas de anotación. Las anotaciones creadas mediante un proceso automatizado debían cargarse en el sistema. Los participantes fueron notificados por correo electrónico y se les presentó un enlace a la interfaz web de la aplicación. Después de iniciar sesión, los anotadores pudían ver una lista de documentos asignados. Se necesitaba una interfaz de usuario intuitiva basada en la web para permitir a los participantes anotar documentos con un mínimo de texto instructivo. El acceso fácil y rápido a las anotaciones fue crucial para el éxito del proyecto. Como el tiempo de los expertos en dominios es bastante valioso, las complicadas instrucciones de instalación o anotación serían prohibitivas. El sistema también necesitaba mostrar estadísticas de acuerdos interannotadores, así como la evaluación.

`Djangology` se puede implantar en cualquier servidor accesible desde la web y requiere una instalación de Python, una instalación de Django y conectividad a un servidor de base de datos. El código fuente y las instrucciones de instalación se pueden encontrar en el sitio web del proyecto[2]. Los autores estiman que el tiempo de instalación y configuración de un extremo a

[2] `http://djangology.sourceforge.net/`

otro para un desarrollador experto en Python y Django es de menos de una hora. Una vez implantado, se puede acceder a la aplicación desde cualquier navegador web; no se necesitan complementos de navegador, instalación de JVM o configuraciones de seguridad personalizadas, ya que la comunicación cliente-servidor se basa en solicitudes HTTP y Ajax estándar.

El esquema de la base de datos de la aplicación y la interfaz de usuario se pueden ampliar y personalizar rápidamente agregando nuevos atributos a la clase de modelo de Python correspondiente. El formulario web correspondiente y el esquema de la base de datos subyacente se actualizan de forma transparente mediante el marco de Django.

La aplicación `Djangology` presenta a los administradores una interfaz para crear o modificar proyectos de anotaciones y administrar usuarios. Los administradores pueden importar documentos (documento único o por lotes) en un proyecto, definir el esquema de anotación del proyecto, crear cuentas de anotadores y asignar anotadores a proyectos específicos y a una lista de documentos. Las anotaciones y los documentos existentes también se pueden cargar fácilmente en el sistema a través de scripts Python personalizados (scripts Django independientes) o mediante una conexión directa a la base de datos `Djangology`. `Djangology` se ha utilizado para importar anotaciones creadas manualmente en formato `Knowtator` y desde `BioScope Corpus` [119], así como anotaciones creadas automáticamente por por las herramientas `GATE` y `UIMA` [120]. En el flujo de trabajo del sistema, a los colaboradores se les suele enviar por correo electrónico la información de autenticación del sistema y se les presenta un enlace a la aplicación.

Una vez que hayan iniciado sesión, los anotadores pueden seleccionar uno de sus documentos asignados y continuar utilizando la interfaz de anotación basada en la web. Una página web basada en Ajax permite a los colaboradores resaltar un fragmento de texto y asignarlo a uno de los tipos de anotaciones predefinidos (según el esquema de anotaciones del proyecto). El procedimiento para ingresar nuevas anotaciones y modificar las existentes es intuitivo y se basa en las convenciones de la interfaz de usuario. El sistema está diseñado específicamente para requerir una inversión mínima de tiempo por parte de los anotadores involucrados. No es necesario instalar, configurar o leer los manuales de usuario por parte de los colaboradores. Las anotaciones se guardan en la base de datos de backend a medida que se ingresan, lo que

garantiza que no se pierda ningún trabajo. Para ahorrar el esfuerzo de los anotadores, una vez que se anota una frase, todas las apariciones de la frase en el documento se anotan automáticamente con la misma etiqueta. Los usuarios también tienen la posibilidad de anular las anotaciones creadas automáticamente o cambiar el comportamiento predeterminado del sistema. Si lo desea, los contribuyentes también pueden marcar los documentos como completados para alertar al administrador del proyecto sobre el progreso de la anotación.

Una vez que se recopilan las anotaciones de varios contribuyentes, los administradores de proyectos tienen la capacidad de ver las estadísticas del acuerdo interannotador: una variedad de métricas basadas en proyectos y documentos por pares se calculan y presentan en la interfaz de usuario. Dado que el análisis de los desacuerdos entre los anotadores es una tarea común, también se proporciona una interfaz para una comparación lado a lado de las anotaciones de los documentos.

MyMiner

`MyMiner` es una aplicación web interactiva basada en un diseño modular con el propósito de ayudar a los usuarios en las tareas de biocuración y anotación de texto. La interfaz de `MyMiner` está diseñada para ser fácil de usar y no requiere la instalación de ningún software local. Cada módulo tiene una opción de exportar para guardar los resultados. El tiempo dedicado a procesar un documento se registra en el archivo exportado. Para mejorar la facilidad de uso, los autores han adoptado y conservado un diseño de pantalla común entre los módulos de la aplicación. El área de análisis del documento de entrada se encuentra en la parte superior de la página; las opciones y herramientas se colocan debajo de la zona de selección principal. `MyMiner` combina PHP, JavaScript y AJAX para mejorar la interactividad del usuario. El núcleo del sistema `MyMiner` cubre cuatro módulos de aplicación que pueden usarse de forma independiente o combinarse siguiendo los pasos de una tubería de biocuración. `MyMiner` maneja cualquier texto sin formato, incluidos resúmenes de artículos, oraciones de documentos, términos de ontología o descripciones de enfermedades.

El módulo "Etiquetado de archivos" es una interfaz de clasificación de texto manual fácil de usar que permite clasificar documentos, resúmenes,

oraciones o términos, ofreciendo la posibilidad de ingresar etiquetas de clase especificadas por el usuario. Este módulo podría usarse, por ejemplo, para clasificar documentos como relevantes o no para un tema específico de una consulta de PubMed. Su propósito es cubrir la tarea de triaje (selección de artículos) que realizan los anotadores de la base de datos, pero también se puede utilizar para cualquier registro de clasificación manual. Los datos etiquetados que resultan de esta clasificación pueden servir como conjuntos de entrenamiento y prueba para sistemas de categorización de texto. Para reducir el tiempo de clasificación manual, el sistema ofrece la opción de establecer dinámicamente fragmentos de texto resaltadoscomo positivos y negativos. Estas son expresiones que los usuarios pueden establecer en cualquier momento durante el proceso de etiquetado para resaltar el texto relevante (marcado en amarillo) o no relevante (marcado en rojo) para el tema de interés. El sistema ofrece la posibilidad de cargar las pautas de clasificación para que el anotador pueda consultarlas cuando sea necesario. Los usuarios pueden pausar y reanudar el proceso de conservación en cualquier momento guardando el documento clasificado. Para reanudar la clasificación, el archivo guardado se carga como archivo de entrada. Se registra el tiempo empleado por un usuario para seleccionar la etiqueta correspondiente. Esto puede resultar útil para estimar la eficiencia de los anotadores y la dificultad de la tarea.

El módulo "Comparar archivo" facilita la comparación directa de colecciones de elementos etiquetados generados por varios enfoques o personas. Además, es posible crear subconjuntos a partir de estas colecciones en función del acuerdo o desacuerdo de las etiquetas de anotación. Este módulo se puede utilizar para comparar y evaluar métodos de clasificación de documentos entre varias personas o softwares. Muestra un resumen global con información que cubre: (i) el número de documentos dentro de cada clase; (ii) el tiempo medio necesario para clasificar el texto; (iii) la correlación entre el tiempo de clasificación y la longitud del texto o (iv) el número de elementos etiquetados de manera diferente entre los anotadores. Este módulo permite extraer una colección de textos (conocido como corpus *Gold Standard*) que han sido etiquetados consistentemente por todos los anotadores. Alternativamente, el módulo también se puede utilizar para extraer los casos límite etiquetados de manera diferente. Las anotaciones apresuradas o inexactas se pueden detectar por desacuerdos entre los anotadores y/o una mala correlación entre

el tamaño del documento y el tiempo de clasificación. Estos casos se pueden usar para refinar y mejorar las pautas de clasificación. El módulo Comparar archivo se ha utilizado para estimar la coherencia de las anotaciones manuales entre varios individuos y métodos.

El módulo "Etiquetado de entidades" (reconocimiento de menciones de entidades) permite detectar manualmente objetos conceptuales importantes dentro de un documento, un primer paso para una mayor identificación de eventos de anotación y relaciones para poblar bases de datos de conocimiento. Este módulo puede usarse, por ejemplo, para crear un corpus de menciones de genes y proteínas para probar y entrenar una herramienta de reconocimiento de entidades nombradas. Ofrece una interfaz interactiva que permite a los usuarios identificar semiautomáticamente varios tipos de entidades dentro de los documentos. Ha sido diseñado como un editor en línea WYSIWYG (What You See Is What You Get) que permite la adición de etiquetas especificadas por el usuario para nuevos tipos de entidades. Para la detección de bioentidades importantes, este módulo proporciona el reconocimiento automático de proteínas, ADN, ARN, líneas celulares y tipos de células mediante la integración del marcador ABNER [121]. `MyMiner` incorpora además el sistema LINNAEUS para identificar especies y organismos [122]. Además, las entidades definidas por el usuario se pueden detectar si se proporcionan diccionarios de términos y etiquetas. Para mejorar la precisión de las anotaciones, las etiquetas se pueden editar y las etiquetas generadas incorrectamente se pueden eliminar.

El módulo "Entity Linking" facilita la anotación manual de bioentidades mencionadas en un documento con identificadores estandarizados. Este módulo puede usarse, por ejemplo, para vincular manualmente artículos a identificadores de enfermedades y proteínas para crear un catálogo de proteínas involucradas en patologías. Los nombres de genes y proteínas se reconocen automáticamente y se muestran como una lista que se puede editar manualmente, y se pueden agregar nuevas entidades y eliminar las identificadas incorrectamente. Para cada nombre de gen o proteína, `MyMiner` sugiere una lista clasificada de identificadores UniProt que utilizan el mecanismo de puntuación de búsqueda UniProt [123]. Para este propósito, `MyMiner` lanza consultas asincrónicas a las respectivas bases de datos (UniProt, taxonomía NCBI, OMIM y archivo de ontología proporcionado por el usuario) utilizando solicitudes AJAX. Para organismos, proteínas, enfermedades y términos de

ontología, se muestra una breve descripción para ayudar a validar posibles aciertos candidatos y para ayudar durante la desambiguación manual de identificadores de bases de datos potenciales. Las casillas de verificación permiten la selección de los identificadores más apropiados de la lista de candidatos. Si las especies se especifican antes de una búsqueda de identificadores de proteínas, se aplican restricciones específicas de especies para reducir el número de candidatos potenciales de UniProt.

WAT-SL

`WAT-SL` es una herramienta de anotación basada en tecnologías web de código abierto dedicada al etiquetado de segmentos. `WAT-SL` proporciona funcionalidades para ejecutar y administrar proyectos de etiquetado de segmentos de manera eficiente. Su interfaz de anotación es autodescriptiva y solo requiere un navegador web, lo que la hace particularmente conveniente para los procesos de anotaciones remotos. La interfaz se puede adaptar a los requisitos del proyecto utilizando tecnologías web estándar. Al mismo tiempo, `WAT-SL` asegura que los textos a etiquetar permanezcan legibles durante todo el proceso de anotación. ste proceso se basa en el servidor y se puede interrumpir en cualquier momento. El progreso del anotador es monitoreable en tiempo real, ya que todas las interacciones relevantes de los anotadores se registran en un archivo de texto sin formato basado en valores clave.

En `WAT-SL`, el proceso de anotación se divide en tareas, que generalmente corresponden a textos individuales y conforman un proyecto de anotación. La interfaz de `WAT-SL` proporciona pistas visuales para ayudar a los anotadores a identificar fácilmente los segmentos anotados. Por ejemplo, para ayudar a los anotadores a formar un modelo mental de la interfaz de anotaciones, los colores de fondo de los segmentos etiquetados coinciden con los colores de las etiquetas en el menú. Todas las etiquetas se guardan automáticamente, evitando pérdida de datos en caso de cortes de energía, problemas de conexión o similares. En algunos casos, los textos pueden estar segmentados en exceso, por ejemplo, debido a una segmentación automática. En estos casos `WAT-SL` permite a los anotadores marcar un segmento para continuar en el siguiente segmento. Adicionalmente, la interfaz de anotaciones incluye un cuadro de texto para introducir comentarios para los organizadores del proyecto. Esto permite una retroalimentación entre los anotadores y los administradores.

Después de que se completa un proceso de anotación, generalmente sigue una fase de curación en la que las anotaciones de diferentes anotadores se consolidan en un resultado final. La interfaz de curación de `WAT-SL` permite una curación eficiente al imitar la interfaz de anotación con tres ajustes. En primer lugar, los segmentos para los cuales la mayoría de los anotadores acordaron una etiqueta están preetiquetados en consecuencia. En segundo lugar, el menú muestra para cada etiqueta cuántos anotadores la eligieron. Finalmente, la descripción de la etiqueta muestra (anonimizada) qué anotador eligió la etiqueta, para que los curadores puedan interpretar cada etiqueta en su contexto.

Desde el punto de vista de la arquitectura de software, `WAT-SL` es una aplicación Java independiente de la plataforma con pocas configuraciones almacenadas en un archivo de configuración llave-valor. Por ejemplo, los anotadores se gestionan en este archivo asignando un nombre de usuario, una contraseña y un conjunto de tareas a cada uno de ellos. Para cada tarea, el organizador de un proyecto de anotación crea un directorio, que `WAT-SL` utiliza como nombre de la tarea en todas las ocasiones.

2.3. Recursos Lingüísticos

En el estado del arte se han establecido diferentes relaciones semánticas, muchas de las cuales dan lugar a la construcción de corpus. Nos enfocamos en dos enfoques: corpus anotados o recursos semánticos para representar el conocimiento en dominios generales, así como aquellos específicamente relacionados con la salud. En esta sección se presenta una comparación entre varios recursos de esta naturaleza así como una breve descripción de cada uno.

La tabla 2.2 presenta diez características relacionadas con la expresividad de los modelos de anotación y los procesos de construcción usados en una muestra de recursos del estado de la técnica. En la Sección 2.3.1 se presenta una descripción más detallada de cada uno de estos recursos. Las características analizadas se pueden entender en los siguientes términos:

Anotación de propósito general Los modelos de anotación de propósito general se utilizan a menudo en corpus extraídos de fuentes enciclopédicas,

Características	**Ixa MedGS**	**DrugSemantics**	**DDI**	**CLEF**	**BARR2**	**Bio AMR**	**EmotiNet**	**YAGO**	**ConceptNet**
propósito general						✓	✓	✓	✓
independencia de la sintaxis	✓	✓	✓	✓			✓	✓	✓
conocimiento ontolológico						✓		✓	✓
conceptos compuestos						✓			
atributos		✓		✓		✓		✓	
relaciones contextuales						✓			
causalidad / implicación	✓			✓		✓			✓
anotación (semi)automática	✓		✓		✓			✓	✓
contenido multi-lingüe				✓	✓		✓		✓
anotadores expertos	✓	✓	✓	✓	✓				

Tabla 2.2: Comparación cualitativa entre recursos lingüísticos del estado de la técnica, tanto de dominio general, como específicos al dominio de la salud.

como `YAGO` [14] y `ConceptNet` [15], los cuales contienen datos extraídos automáticamente de Wikipedia. (entre otras fuentes). Por el contrario, los modelos de anotaciones de dominios específicos se suelen emplear cuando la fuente está más restringida a un dominio específico. Los ejemplos incluyen `Ixa MedGS` [10], que contiene conceptos relacionados con la salud para enfermedades, causas y medicamentos; `DrugSemantics` [11], que anota entidades de salud, medicamentos y procedimientos; y `DDI` [12], que anota las interacciones fármaco-fármaco. Un término medio es el corpus `Bio AMR` [13], que aplica un modelo de anotación de propósito general (`AMR`) [9] a los documentos de salud.

La mayoría de los recursos mencionados se centran en capturar la semántica de las oraciones, en el sentido de que es probable que se anoten de manera similar oraciones muy diferentes con los mismos hechos. Consideramos que `BioAMR` es menos independiente de la sintaxis porque aunque AMR es un modelo de anotación semántica — mucho más abstracto que el análisis de dependencia, por ejemplo —, todavía se basa en gran medida en la estruc-

tura gramatical de las oraciones. Por lo tanto, es probable que un cambio significativo en la estructura de la oración cambie la anotación, incluso si el mensaje semántico subyacente permanece sin cambios. Por ejemplo, dado que `AMR` usa roles `PropBank` [16], cambiar una palabra por una palabra semánticamente similar, incluido un sinónimo, probablemente cambiará la anotación correspondiente y, por lo tanto, los roles disponibles. Esto también hace que `AMR` y recursos similares dependan del idioma, no solo en la práctica dada su dependencia de la existencia de bancos de palabras, sino desde su concepción. Al intentar aplicar `AMR` en español, Migueles-Abraira et al. [17] muestra que aunque `AMR` es teóricamente independiente del idioma, las pautas de anotaciones existentes están sesgadas hacia el inglés y deben adaptarse para capturar fenómenos lingüísticos que no existen en inglés.

En contraste, la base de conocimiento de `EmotiNet` [33] está orientada hacia un dominio específico (emociones), y se construye a partir de la anotación manual de entradas de blog, utilizando una estructura semántica general que vincula entidades, acciones y emociones. Aunque `EmotiNet` está diseñado para un dominio en particular, su estructura es bastante general, en el sentido de que puede representar fácilmente cualquier tipo de evento o acción realizada por entidades.

Conocimiento ontológico Los modelos de anotación de propósito general a menudo permiten representar el conocimiento ontológico en forma de herencia y composición entre conceptos. En este contexto, consideramos la capacidad de reconocer y anotar estas relaciones ontológicas en el texto fuente. Los modelos de anotación relacionados con la salud no suelen abordar este problema, principalmente porque las entidades y relaciones a anotar forman una ontología predefinida donde la composición y jerarquía, si existe, ya están concebidas en el propio modelo de anotación. Sin embargo, las anotaciones de propósito general a menudo incluyen relaciones como *is-a* o *part-of* que representan directamente estos conceptos ontológicos y, por lo tanto, pueden extraer representaciones ontológicas del texto natural.

Conceptos compuestos Los conceptos compuestos, por el contrario, se refieren a la capacidad de anotar conceptos que están formados por una combinación fina de otras entidades, en la misma oración. Por ejemplo, tome la oración: " *a los médicos que trabajan en el turno de noche se les pagan*

horas extra ". AMR permite la representación del concepto de que no todos los médicos, sino solo aquellos que trabajan en el turno de noche, son los que cobran horas extra.

Atributos Los atributos se utilizan a menudo para refinar aún más el significado de las entidades anotadas. Los ejemplos incluyen cuantificadores en AMR o modificadores que especifican un grado de incertidumbre o una negación de un concepto.

Relaciones contextuales Las relaciones contextuales permiten presentar hechos que solo ocurren bajo ciertas condiciones, por ejemplo, en un marco de tiempo o lugar específico o bajo ciertos supuestos. Esto permite una anotación semántica más detallada. BioAMR hereda esta habilidad de AMR, que permite modificadores para expresar *cómo*, *cuándo*, *dónde* o *por qué* ocurre algún evento.

Causalidad y vinculación La causalidad y la implicación son relaciones de propósito general que permiten cierto nivel de inferencia o razonamiento. El corpus Ixa MedGS define una relación *causa*, ya que es relevante en el dominio que el corpus está modelando. Asimismo, AMR y ConceptNet incluyen relaciones similares.

Anotación (semi)automática Otra característica interesante es el tipo de anotación, ya sea manual, pre-automatizada con revisión de expertos o totalmente automatizada. Aunque investigaciones recientes muestran una tendencia creciente hacia la anotación pre-automatizada o totalmente automatizada, la anotación manual todavía se considera más confiable. La anotación manual todavía se considera más confiable. Un ejemplo es el corpus DDI [12], que contiene 1025 documentos en inglés de Medline, fue previamente anotado automáticamente y luego revisado manualmente por expertos del dominio (farmacéuticos).

De forma similar, el corpus de Ixa MedGS [124] fue previamente anotado automáticamente y luego revisado manualmente por expertos del dominio en farmacología. En general, los recursos semánticos de mayor volumen, como ConceptNet y YAGO, se diseñan para ser anotables de forma automática o

semi-automática, a partir de reglas de extracción diseñadas por expertos del dominio, o obtenidas indirectamente por sistemas de aprendizaje automático.

Contenido multi-lingüe Algunos recursos lingüísticos contienen contenido en más de un idioma. Esta característica requiere que el modelo de anotación subyacente sea relativamente independiente de la sintaxis, y que, o bien existan anotadores capaces de reconocer diferentes idiomas, o se emplee alguna tecnología de extracción automática.

Anotadores expertos En función de la complejidad del modelo de anotación y del contenido textual, es posible que se requiera de cierto nivel de experticia técnica en el dominio a anotar. Los corpus relacionados con la salud suelen ser anotados por expertos con una estructura semántica de dominio específico, como entidades relacionadas con enfermedades, fármacos, genes o tratamientos. Dada la complejidad de los conceptos en el dominio médico, los anotadores suelen incluir médicos u otros especialistas del dominio médico. En estos recursos, se utilizan muy pocas características del lenguaje natural de propósito general. Esto proporciona un mayor detalle de la información semántica, ya que las entidades y relaciones son relevantes para el dominio en cuestión. Sin embargo, en el mismo sentido, podría descartar información importante en el texto que no se puede representar con la estructura definida. Por el contrario, los corpus o bases de conocimiento de propósito general suelen ser anotados por no expertos con una estructura semántica diseñada para representar tanto conocimiento como sea posible. Esta estrategia tiende a aumentar la memoria (se extrae una mayor cantidad de hechos) pero puede extraer hechos irrelevantes o incorrectos.

2.3.1. Descripción de los Recursos Lingüísticos

A continuación se presenta una breve descripción de cada uno de los recursos lingüísticos considerados en esta Tesis.

Ixa MedGS

Consiste en resúmenes de alta anotados sintáctica y semánticamente escritos en idioma españolpor personal médico. Este corpus fue diseñado para

desarrollar herramientas de anotación automática y, por lo tanto, facilitar a los médicos la recuperación de información de las historias clínicas. Los datos consisten en historias clínicas electrónicas recogidas en el Hospital Galdakao-Usansolo. Fueron recopilados a partir de un convenio entre el Servicio Vasco de Salud y la Universidad del País Vasco, en el que el Servicio de Salud facilitó un corpus anonimizado eliminando los datos identificativos de carácter personal y autorizó su uso exclusivamente con fines de investigación.

El corpus se centra en la anotación de los *efectos adversos* definidos en un informe (*Estudio Nacional sobre Efectos Adversos asociados a la hospitalización*) publicado por el Ministerio de Sanidad y Consumo español como todo accidente o incidente que haya herido o pueda haber lesionado al paciente durante el tratamiento. Entre los diferentes efectos adversos que se distinguen en el informe, la gran mayoría está relacionada con una de las siguientes tres causas: (i) prescripción de medicamentos (37,4 % de todos los EA); (ii) infecciones nosocomiales (25,3 % de todos los EA); y (iii) procedimientos (25,0 % del total de EA). El corpus `IxA MedGS` se concentra especialmente en las *reacciones adversas a los medicamentos*, definidas como trastornos inevitables o difíciles de evitar, con o sin lesión, que se producen cuando los medicamentos se utilizan de forma adecuada.

DrugSemantics

`DrugSemantics` es una colección de resúmenes en español de las características de diferentes productos farmaceúticos. Se anotó manualmente con entidades con nombre farmacoterapéutico, que se detallan en el esquema de anotaciones de `DrugSemantics`. Los anotadores fueron una Enfermera Registrada y dos estudiantes de la Licenciatura en Enfermería. La calidad del corpus de `DrugSemantics` se evaluó midiendo su fiabilidad de anotación (F general = 79,33 % [IC 95 %: 78,35–80,31]), así como su precisión de anotación (global [IC 95 %: 94,11–95,19]). En total, el corpus contiene más de 2 000 entidades con nombre, 780 oraciones y 226 729 palabras. Por último, los autores presentan un módulo de Clasificación de Entidades Nombradas entrenado en `DrugSemantics` con el objetivo de mostrar la calidad del corpus, así como un ejemplo de cómo usarlo. Esto muestra que el copus anotado puede utilizarse para crear algortimos de anotación automática.

DDI

DDI (*drug-drug interaction*) es un corpus anotado con sustancias farmacológicas así como las interacciones entre ellas. Este corpus incluye interacciones entre medicamentos tanto farmacodinámicas como farmacocinéticas. Una interacción farmacodinámica ocurre cuando los efectos farmacológicos de un fármaco se modifican por la presencia de otro fármaco, mientras que una farmacocinética es el resultado de la interferencia de la absorción, distribución, metabolismo y / o eliminación de un fármaco por otro fármaco. La motivación para crear este corpus radica en la escasez de recursos anotados para la extracción de interacciones farmacológicas, que es uno de los principales obstáculos en el desarrollo de sistemas de procesamiento de lenguaje natural, para esta área de farmacovigilancia.

El corpus DDI fue desarrollado para el desafío DDI Extraction 2013[3], cuyo objetivo principal fue proporcionar un marco común para la evaluación de técnicas de extracción de información aplicadas al reconocimiento de sustancias farmacológicas y la detección de DDI a partir de textos biomédicos. En este desafío se propusieron dos etapas: el reconocimiento y clasificación de nombres de fármacos y la extracción y clasificación de sus interacciones.

Las fuentes de contenido textual para la construcción del corpus DDI fueron DrugBank [] y Medline [], resultando en dos colecciones, DDI-DrugBank, y DDI-Medline, respectivamente. Basado en la división de oraciones durante el preprocesamiento, el corpus DDI-DrugBank contiene 6795 oraciones, y el corpus DDI-MedLine se compone de 2147 oraciones. El tipo de entidad más común presente en ambas colecciones fueron los *fármacos* (63 %). Sin embargo, los números de otros tipos de entidades difieren entre ambas colecciones. En cuanto a las relaciones, el *efecto* fue la relación dominante encontrada en todo el corpus DDI.

CLEF

El corpus CLEF consta de registros estructurados y documentos de texto libre de 20234 pacientes, provenientes del Royal Marsden Hospital (RMH), uno de los centros oncológicos especializados más grandes de Europa. Los documentos de texto libre constan de tres tipos: narrativas clínicas; informes

[3] http://www.cs.york.ac.uk/semeval-2013/task9/

de histopatología; e informes de imágenes. La confidencialidad de los pacientes fue garantizada mediante una variedad de medidas técnicas y organizativas, incluida la seudonimización automática y la inspección manual.

Dado el costo de la anotación humana, la porción de referencia del corpus consiste en un subconjunto relativamente pequeño de todo el corpus de 565000 documentos. Para evitar eventos que son raros o que están fuera de los requisitos principales del proyecto, está restringido por diagnóstico y solo considera documentos de aquellos pacientes con un código de diagnóstico primario en una de las subcategorías de nivel superior del Capítulo II de la CIE-10 (neoplasias). Además, solo contiene aquellas subcategorías que cubren más del 5 % de narrativas e informes. El corpus de referencia consta de dos porciones, seleccionadas para propósitos ligeramente diferentes.

Registros completos de pacientes Dos aplicaciones en `CLEF` implican agregar datos en un solo registro de paciente. Estas dos aplicaciones requieren registros completos de pacientes para su desarrollo y prueba. Para ello, se seleccionaron dos registros completos de pacientes para esta parte del corpus, de dos de las principales categorías de diagnóstico, para dar un número medio de documentos y una combinación de tipos y longitudes de documentos.

Muestra aleatoria estratificada La mayor parte del corpus de referencia sirve como material de desarrollo y evaluación para tareas de extración de información. Con el fin de garantizar un entrenamiento uniforme y una evaluación justa en todo el corpus, el muestreo de esta porción es aleatorio y estratificado, de modo que refleje la distribución de los documentos. La anotación inicial de la muestra aleatoria se centra en 50 de cada uno de los relatos clínicos, los informes de histopatología y los informes de imágenes, para un total de 150 documentos anotados.

El corpus de referencia `CLEF` es un corpus anotado semánticamente, concentrado en las principales entidades semánticas del texto. Se considera entidad a cualquier elemento del mundo real a la que se hace referencia en el texto: los fármacos que se mencionan, las pruebas que se realizaron, etc. Además, se anotan las relaciones entre entidades: la condición indicada por un fármaco, el resultado de una investigación, etc.

La anotación está anclada en el texto. Los anotadores marcan tramos de

texto con un tipo: fármaco, lugar geométrico, etc. Los anotadores también pueden marcar palabras que modifican tramos (como negación) y marcar relaciones como vínculos entre tramos. Dos o más intervalos pueden hacer referencia al mismo elemento en el mundo real, en cuyo caso son co-referencia.

Los tipos de anotaciones se describen en un esquema basado en un conjunto de requisitos desarrollados entre médicos y lingüistas computacionales en `CLEF` (no existen esquemas o teorías estándar en esta área). Los tipos de esquema se asignan a tipos en la red semántica UMLS. A los efectos de la anotación, el esquema se modela como una ontología. La anotación se lleva a cabo utilizando una versión adaptada de la herramienta `Knowtator` (ver Sección 2.2.2).

La metodología de anotación sigue los estándares establecidos de procesamiento del lenguaje natural.Todos los documentos están anotados por al menos dos anotadores, y solo se utilizan cuando el acuerdo supera un umbral de consenso. En tal caso, las diferencias las resuelve un tercer anotador experimentado.

Dado que la coherencia es fundamental para la calidad de un corpus de referencia, es importante que todos los documentos estén anotados con el mismo estándar. Para garantizar la coherencia, en `CLEF` se proporcionó un conjunto de pautas a los anotadores. Estas describen en detalle lo que debe y no debe anotarse; cómo decidir si dos entidades están relacionadas; cómo lidiar con la co-referencia; y varios casos especiales. Las pautas también proporcionan una secuencia de pasos que los anotadores deben seguir al trabajar en un documento, diseñada para minimizar los errores de omisión.

La concordancia entre documentos con anotaciones dobles se mide en `CLEF` utilizando la concordancia entre anotadores. Los pares de anotaciones dobles se rechazan si el acuerdo no supera un umbral preestablecido. El proceso de resolución lo lleva a cabo un tercer anotador experimentado. Todos los acuerdos de los anotadores originales se aceptan en un conjunto de consenso, y el tercer anotador decide sobre las diferencias, de acuerdo con un conjunto de pautas estrictas.

Más allá del corpus anotado, `CLEF` propone una metodología para la anotación de un estándar de oro para la extracción de información clínica y demuestra que es viable. Los resultados iniciales muestran que se pueden lograr niveles prometedores de acuerdo entre los anotadores, aún aquellos

sin experticia técnica en el dominio del conocimiento concreto que se está anotando.

BARR2

BARR2 es un corpus anotado manualmente seleccionando 684 casos clínicos de SciELO España, distribuídos en 318 casos clínicos para el conjunto de entrenamiento, 146 para el conjunto de desarrollo y 220 para el conjunto de prueba. El etiquetado manual de las menciones de abreviaturas del corpus se realizó utilizando una versión personalizada de Annotator. Luego, se utilizó el kit de herramientas de anotación Brat para revisar manualmente las anotaciones de menciones y anotar las relaciones entre las formas cortas y sus correspondientes formas largas, así como las menciones anidadas.

El corpus creado está compuesto por diferentes casos clínicos de acceso abierto redactados en español. Los expertos en el dominio anotaron las abreviaturas presentes en estos informes, junto con su definición. Los resultados del consenso entre los anotadores mostraron que los anotadores estuvieron de acuerdo con el 88 % de los casos etiquetados, y concluyeron que las directrices que utilizaron estaban claramente estructuradas. Examinando cuáles son los tipos de abreviaturas más frecuentes, se puede observar que muchas de ellas corresponden a unidades de medida (clave para la detección de posología y posología), entidades anatómicas, marcadores bioquímicos y tratamientos. Al observar el idioma de las definiciones de abreviaturas, solo alrededor del 68 % por ciento correspondió a definiciones en español, el resto donde la mayoría de las definiciones en inglés a menudo se relacionan con sustancias, tratamientos y entidades bioquímicas.

Para que el proceso de anotación sea lo más flexible posible, los anotadores tuvieron acceso al artículo completo del caso clínico para encontrar sus significados mencionados explícitamente en otras secciones. Aemás, los anotadores utilizaron varios diccionarios biomédicos españoles de abreviaturas y acrónimos. Previo a la anotación, se aplicó un algoritmo automático de detección de pares de abreviaturas-definiciones para detectar diferentes pares en el texto completo, por lo que los anotadores no necesitaban leer todo el artículo buscando definiciones explícitas, a menos que fuera necesario. Finalmente, se empleó una base de datos interna de pares de abreviaturas y definiciones en español para asignar definiciones a las abreviaturas sin

definiciones explícitas.

Bio AMR

El corpus `Bio AMR` fue construido para la tarea evaluativa de reconocimiento y generación de AMR en el evento Semeval 2017 []. Incluye 6452 oraciones anotadas en format `AMR` de varias fuentes relacionadas con el cáncer extraídas de PubMed. Entre ellas, se anotaron tres artículos completos de esta colección, así como las secciones de resultados de 46 artículos adcionales de la misma base de datos. El corpus también incluye alrededor de 1000 oraciones, cada una del corpus de entrenamiento BEL BioCreative y del Chicago Corpus.

EmotiNet

`EmotiNet` es una base de conocimientos para almacenar cadenas de acción y sus correspondientes etiquetas emocionales de diversas situaciones. Una cadena de acción se define en `EmotiNet` como una secuencia de vínculos de acción, o simplemente acciones que desencadenan una emoción en un actor. Cada vínculo de acción específico se puede describir con una tupla *(actor, acción, paciente, emoción)*.

El propósito de `EmotiNet` es modelar situaciones como cadenas de acciones y su correspondiente efecto emocional utilizando una representación ontológica. El conocimiento representado en `EmotiNet` tiene el objetivo de ser compartido por la comunidad científica. Así mismo, este conocimiento debe ser alimentado por fuentes heterogéneas de conocimiento común para evitar incertidumbres. Sin embargo, se pueden introducir afirmaciones específicas para dar cuenta de las especificidades de los individuos o contextos. Estas características permiten modelar la interacción de diferentes eventos en el contexto en el que tienen lugar y agregar mecanismos de inferencia para extraer conocimiento que no está explícitamente presente en el texto. Además, permite incluir conocimientos sobre los criterios de valoración relacionados con diferentes conceptos encontrados en otras ontologías y bases de conocimiento (para dar cuenta de las diferentes propiedades del actor, acción y objeto).

YAGO

YAGO surge por la necesidad de crear una ontología más grande utilizando las ontologías existentes actualmente, sus principales objetivos eran:

- Unificación de Wikipedia y WordNet.
- Hacer uso de estructuras e información, tales como: Infoboxes, Category Pages, etc.
- Asegurar la plausibilidad de los hechos a través de comprobación de tipo

La entrada de datos de YAGO se compone de cuadros de información semiestructurada de artículos de Wikipedia y entradas de WordNet y Geonames(en versiones posteriores). El resultado es una base de conocimiento que en su primera versión en 2007 contaba con más de 1 millón de entidades y 5 millones de hechos. En versiones posteriores se ha extendido a 350K tipos de entidades, 10M entidades, 120M hechos e incorporando información temporal y espacial.

La estructura del modelo de representación de YAGO es una extensión de RDFS que incluye transitividad acíclica. Las entidades son objetos ontológicos abstractos. Cada entidad es parte de al menos una clase y las clases están ordenadas en jerarquía taxonómica. Las relaciones son un tipo especial de entidad. Los hechos son tripletas de la forma: entidad, relación, entidad. Cada hecho tiene un identificador único.

La contribución clave de YAGO es que proporciona una vinculación semántica de categorías de Wikipedia a Wordnet, con una alta calidad en la extracción de conocimiento (la precisión de los hechos extraídos es estimada en un 95 %). Además, la estructura del conocimiento representado en YAGO es decidible, extensible, y compatible con RDFS.

ConceptNet

ConceptNet es un proyecto de representación del conocimiento, que proporciona un gran grafo semántico que describe el conocimiento humano general y cómo se expresa en lenguaje natural. El alcance de ConceptNet incluye palabras y frases comunes en cualquier lenguaje humano escrito.

Proporciona un amplio conjunto de conocimientos previos que debe conocer una aplicación informática que trabaja con texto en lenguaje natural. Estas palabras y frases se relacionan a través de un dominio abierto de predicados, que describen no solo cómo se relacionan las palabras por sus definiciones léxicas, sino también cómo se relacionan a través del conocimiento común. Por ejemplo, su conocimiento sobre el "jazz" incluye no solo las propiedades que lo definen, como `IsA (jazz, music_genre)`, sinn también incluye hechos incidentales como *AtLocation (jazz, new_orleans)* y *UserFor (saxo, jazz)*.

`ConceptNet` se originó como una representación del conocimiento recopilado por el proyecto Open Mind Common Sense [125], que utiliza un sitio web interactivo de larga duración para recopilar nuevas declaraciones de los visitantes del sitio y les hace preguntas específicas sobre afirmaciones que cree que pueden ser ciertas. Los lanzamientos posteriores incluyeron conocimientos de sitios web similares en otros idiomas, como portugués y holandés, y colaboraciones con juegos de palabras en línea que recopilan automáticamente conocimientos generales, lo que proporciona más conocimientos en inglés, japonés y chino.

`ConceptNet` proporciona una base de conocimiento del mundo real para una variedad de proyectos y aplicaciones de Inteligencia Artificial. Se han utilizado versiones anteriores de `ConceptNet`, por ejemplo, para construir un sistema para analizar el contenido emocional del texto [126], para crear un sistema de diálogo para mejorar las especificaciones del software [127], y para visualizar temas y tendencias en un corpus de texto no estructurado [128].

`ConceptNet` proporciona una combinación de características que no están disponibles en otros proyectos de representación del conocimiento:

- Sus conceptos están conectados a palabras y frases en lenguaje natural que también se pueden encontrar en texto libre.
- Incluye no solo definiciones y relaciones léxicas, sino también las asociaciones de sentido común que la gente común hace entre estos conceptos. Sus fuentes varían en formalidad desde diccionarios hasta juegos en línea.
- Los conceptos no se limitan a un solo idioma; pueden ser de cualquier idioma escrito.
- Integra conocimientos de fuentes con distintos niveles de granularidad

y distintos registros de formalidad, y los pone a disposición a través de una representación común.

`ConceptNet` tiene como objetivo contener tanto hechos específicos como el desordenado e inconsistente mundo del conocimiento de sentido común. Para comprender verdaderamente los conceptos que aparecen en un texto en lenguaje natural, es importante reconocer las relaciones informales entre estos conceptos que forman parte del conocimiento cotidiano, que a menudo están infrarrepresentados en otros recursos léxicos. `WordNet`, por ejemplo, puede representar que un perro es un tipo de carnívoro, pero no que es un tipo de mascota, o que un tenedor es un utensilio para comer, pero no tiene ningún vínculo entre *tenedor* y *comer* para indicar que se usa un tenedor para comer.

Agregar conocimiento de sentido común crea muchas preguntas nuevas, cómo decir que "un tenedor se usa para comer" si un tenedor se usa para otras cosas además de comer, o distinguir el utensilio para comer de la acepción que indica una bifurcación. `ConceptNet` intentar recopilar representaciones que respondan a estas preguntas, mientras aceptamos pragmáticamente que gran parte del contenido de una base de conocimiento de sentido común las dejará sin resolver.

2.4. Aprendizaje Automático en el Descubrimiento de Conocimiento

En los últimos años, investigadores en campos como el aprendizaje automático, el descubrimiento de conocimientos, la minería de datos y el procesamiento del lenguaje natural, entre otros, han producido muchos enfoques y técnicas para aprovechar la gran cantidad de información disponible en Internet para una variedad de tareas, desde la creación de búsquedas. [129] y sistemas de recomendación [130] hasta para mejorar los diagnósticos médicos [131].

Entre los diferentes enfoques relevantes para el descubrimiento del conocimiento, podemos reconocer un espectro continuo de técnicas, basado en cuánto conocimiento experto se utiliza. Las técnicas fundamentalmente basadas en el conocimiento consisten en reglas definidas en bases de conocimiento

elaboradas a mano por expertos en el dominio [132]. Estos enfoques tienen un alto grado de confiabilidad y precisión, y generalmente permiten una mayor complejidad en el conocimiento extraído, pero son difíciles de escalar a grandes cantidades de datos. Por el contrario, los enfoques estadísticos consisten en técnicas basadas en el reconocimiento de patrones con modelos estadísticos y probabilísticos [133]. Estas técnicas escalan mejor con grandes cantidades de datos [134], proporcionando una mejor recuperación, pero a menudo se limitan a extraer modelos simples de conocimiento y pueden ser más sensibles a información ruidosa, falsa o sesgada [135].

Dadas estas características mutuamente complementarias, se han propuesto varios enfoques híbridos. Recientemente, han surgido áreas de investigación como el aprendizaje de ontología [52], el aprendizaje mediante la lectura [53] o la incorporación de entidades [136]. En estas áreas, los investigadores combinan técnicas del aprendizaje automático, el procesamiento del lenguaje natural y la representación del conocimiento para resolver problemas más complejos que no se pueden abordar utilizando solo las herramientas clásicas.

Muchos sistemas de aprendizaje automático están diseñados para resolver una tarea específica de un dominio, como asignar una clase a un elemento de un conjunto predefinido de etiquetas. Estos sistemas, cuando se entrenan con datos para un dominio en particular, a menudo no son aplicables a otros dominios o a escenarios donde varios dominios diferentes deben utilizarse juntos. Además, a menudo los sistemas están diseñados para ser entrenados una vez a partir de un corpus y no permiten una mejora continua del conocimiento aprendido. Recientemente, hay intentos de construir sistemas de aprendizaje de propósito general que siempre mejoran mientras obtienen nuevos conocimientos, reevalúan los conocimientos antiguos y refinan su propia confianza [32].

Además, es interesante diseñar sistemas de aprendizaje no monolíticos, sino que se construyen como un conjunto de componentes modulares que se pueden combinar de diferentes formas. Esta componibilidad permitiría un sistema de aprendizaje continuo no solo para mejorar la calidad del conocimiento extraído, sino también para aprender a sintonizar sus propios parámetros internos para realizar una mejor extracción de conocimiento en el futuro. Es concebible que dicho sistema pueda aprender gradualmente qué tipos de procesos básicos (es decir, reconocimiento de entidades, etiquetado

POS, etc.) son más útiles para un dominio dado o para un corpus en particular. Asimismo, dicho sistema podría aprender qué tipos de modelos probabilísticos proporcionan los mejores resultados en un conjunto de datos en particular.

La velocidad y el volumen de producción de la información se ha incrementado exponencialmente en la última década, principalmente por el auge de las redes sociales y la tecnología móvil. Para hacer frente a este volumen de información, es necesario poder procesar cantidades masivas de datos de forma continua. El campo del aprendizaje automático proporciona herramientas para la extracción automática de información y conocimiento de diferentes fuentes de datos. El aprendizaje automático no solo permite automatizar procesos y tareas de descubrimiento de conocimiento o minería de textos, sino que también proporciona una gran mejora en la escalabilidad de estos procesos [26]. Mediante el uso de recursos informáticos masivos, es posible procesar millones de documentos sin procesar en un tiempo razonable, superando con creces lo que pueden hacer los expertos en el dominio. Las mejoras recientes en las capacidades informáticas y el acceso a conjuntos de datos más grandes han dado lugar al campo del aprendizaje profundo, que ha mejorado el estado del arte en varias de las tareas clásicas de aprendizaje automático [137].

Podría decirse que los dos enfoques más comunes en el aprendizaje automático son el aprendizaje supervisado y no supervisado [133]. El aprendizaje supervisado se puede utilizar para reconocer elementos específicos de conocimiento en una fuente de datos. Por ejemplo, etiquetar fragmentos de texto para indicar que definen una entidad [138] (p. Ej., Una persona, organización o lugar), reconocer relaciones entre dichas entidades o asignar un sentimiento o puntuación de opinión [139] a un fragmento de texto. Por otro lado, el aprendizaje no supervisado puede ayudar a encontrar la estructura relevante en un gran conjunto de elementos. Los algoritmos de agrupación en clústeres se pueden utilizar para detectar conceptos similares o para extraer conceptos abstractos de grupos de elementos más concretos. Se pueden utilizar otras técnicas para reducir la cantidad de información, por ejemplo, para eliminar piezas de información ruidosas, inciertas o irrelevantes [140].

En general, la mayoría de los algoritmos de aprendizaje automático no están diseñados para representar el conocimiento aprendido en estructuras complejas, como las definidas por expertos en el dominio humano (es decir,

ontologías). A su vez, las representaciones suelen tener una estructura simple, como una distribución de probabilidad o una matriz de correlación [141]. Al aplicar estos algoritmos a un problema real, se debe realizar una interpretación específica de dominio de esas representaciones.

Además, muchos de los modelos de aprendizaje automático más poderosos son difíciles de explicar, en el sentido de que cuando el sistema produce una respuesta, un experto humano no puede comprender y reproducir fácilmente los pasos de inferencia que realiza el sistema [142]. La elección de una representación adecuada es decisiva para el éxito de la mayoría de las técnicas de aprendizaje automático [143]. En los últimos años, ha aumentado el interés en el problema del aprendizaje automático de representaciones relevantes. Los *embeddings* de palabras [144] y los *embeddings* de entidades más generales [136] representan los primeros pasos para impulsar los enfoques de aprendizaje profundo con representaciones internas más explicables. Dado que las ontologías son, por definición, representaciones de una conceptualización determinada, es concebible que utilizando ontologías como semillas para la representación de un dominio determinado, se pueda mejorar el rendimiento de los procesos de minería de datos basados en el aprendizaje automático.

2.4.1. Anotación Semi-Automática con Aprendizaje Activo

El aprendizaje automático, y específicamente el aprendizaje supervisado, es una de las herramientas más efectivas para automatizar tareas cognitivas complejas, como reconocer objetos en imágenes o comprender texto en lenguaje natural. Uno de los principales obstáculos del aprendizaje supervisado es la necesidad de conjuntos de datos de alta calidad de muestras etiquetadas en las que se puedan entrenar modelos estadísticos. Estos conjuntos de datos generalmente son construidos por expertos humanos en un proceso manual largo y costoso. El aprendizaje activo [145] es un paradigma alternativo al aprendizaje supervisado convencional que se ha propuesto para reducir los costos involucrados en la anotación manual.

La idea clave que subyace al aprendizaje activo es que un algoritmo de aprendizaje puede funcionar mejor con menos ejemplos de entrenamiento si se le permite seleccionar activamente qué ejemplos aprender de [146]. En el contexto del aprendizaje supervisado, este paradigma cambia el papel del

experto humano. En contextos de aprendizaje supervisado convencionales, el experto humano guía el proceso de aprendizaje proporcionando un gran conjunto de datos de ejemplos etiquetados. Sin embargo, en el aprendizaje activo, el rol activo se traslada al algoritmo y el experto humano se convierte en un oráculo, participando en un ciclo de etiquetado, entrenamiento, y consulta. En el paradigma activo, un modelo de aprendizaje automático se construye de forma incremental entrenando en una colección parcial de muestras y luego seleccionando una o más muestras sin etiquetar para consultar el oráculo humano en busca de etiquetas y aumentar el conjunto de entrenamiento. Este enfoque introduce el nuevo problema de cómo seleccionar mejor las muestras de consulta para maximizar el rendimiento del modelo y minimizar el esfuerzo del participante humano.

El escenario de aprendizaje activo más simple consiste en la clasificación de elementos independientes x_i extraídos de un grupo de muestras sin etiquetar. Los ejemplos van desde la clasificación de imágenes [147] hasta la minería de sentimientos [148], en la que el nivel mínimo de muestreo (por ejemplo, una imagen o un documento de texto) corresponde al nivel mínimo de decisión (es decir, se asigna una sola etiqueta a cada x_i). Los escenarios más complejos surgen cuando el nivel de decisión es más detallado que el nivel de muestreo. En el dominio de la minería de texto, un escenario interesante es la tarea de extracción de entidades y relaciones del texto en lenguaje natural [149]. En este escenario, el nivel de muestreo es una oración, pero el nivel mínimo de decisión involucra cada token o par de tokens en la oración y, además, estas decisiones en general no son independientes dentro de la misma oración. En este caso, no es trivial estimar qué tan informativa será una muestra sin etiquetar, ya que cada muestra tiene varias fuentes de incertidumbre.

En esta sección se revisan algunas de las investigaciones más relevantes relacionadas con el aprendizaje activo en general, y específicamente enfocadas en la detección de entidades y extracción de relaciones. Una de las decisiones de diseño más importantes en el aprendizaje activo es cómo seleccionar inteligentemente las nuevas muestras sin etiquetar de la manera más eficiente. El supuesto subyacente es que se desea entrenar un modelo con el rendimiento más alto posible (medido en precisión, F_1, etc.) mientras se minimiza el costo humano (medido en tiempo, número de muestras etiquetadas manualmente o cualquier otra métrica adecuada). Este requisito a menudo se enmarca como la selección de las muestras sin etiqueta *más informativas* y se formaliza en

términos de una *estrategia de consulta* [146]. Las estrategias de consulta más comunes para el aprendizaje activo de propósito general se pueden agrupar en las siguientes categorías:

(i) **Muestreo de incertidumbre:** Las muestras más informativas se consideran aquellas con el mayor grado de incertidumbre, dada alguna medida de incertidumbre para cada muestra [150].

(ii) **Consulta por comité:** Las muestras más informativas se consideran aquellas con el mayor desacuerdo entre un comité de diferentes modelos o diferentes hipótesis del mismo modelo subyacente [146].

(iii) **Cambio de modelo esperado:** Las muestras más informativas se consideran aquellas que producirían el mayor cambio en la hipótesis del modelo si se incluyeran en el conjunto de entrenamiento [151].

(iv) **Reducción de varianza y error:** Las muestras más informativas son aquellas que producen la mayor reducción en el error de generalización del modelo o, como proxy, su varianza [152].

Las estrategias de cambio de modelo (iii) y reducción de varianza / error (iv) esperadas dependen en gran medida del modelo de aprendizaje específico utilizado. Por el contrario, el muestreo de incertidumbre (i) y la consulta por comité (ii) son aplicables en general con un alto grado de agnosticismo del modelo. Además, los subconjuntos relevantes de ambas estrategias se pueden formalizar bajo un solo marco si definimos la incertidumbre como una medida de la entropía de la salida prevista del modelo. En este marco, la consulta por comité se puede implementar mediante votación ponderada, asignando así probabilidades empíricas a los posibles resultados.

La *densidad ponderada* es una estrategia complementaria en la que las muestras más informativas se ponderan por lo representativas que son del espacio de entrada, por ejemplo, midiendo su similitud con las muestras restantes [153]. Este enfoque intenta contrarrestar una tendencia notable a seleccionar ejemplos atípicos como las muestras más informativas, un problema asociado con otras estrategias de consulta, ya que los ejemplos atípicos son a menudo las muestras que crean la mayor cantidad de incertidumbre, desacuerdo o cambio de hipótesis.

Los avances recientes en el procesamiento de lenguaje natural han producido un crecimiento del interés en técnicas de aprendizaje activo para alivar los requisitos y costos de anotación de corpus largos [154, 155]. Settles and Craven [153] compara varias estrategias para aprendizaje activo en tareas de etiquetado de secuencias, concluyendo que las estrategias de consulta basadas en medidas de entropía de las secuencias combinadas con un muestro ponderado por densidad tienen un mejor rendimiento que las estrategias alternativas. Meduri et al. [156] propone un entorno de evaluación para diferentes técnicas de aprendizaje activo en el contexto de tareas de normalización y enlazado de entidades. En la tarea de reconocimiento de entidades nombradas, se han utilizado modelos CRF para seleccionar los ejemplos de consulta [157, 158]. Por otro lado, la tarea de extracción de relaciones también se ha beneficiado de enfoques de aprendizaje activo, tanto en escenarios de propósito general [159] como en dominios específicos [149]. Sin embargo, a pesar del creciente interés en esta área de investigación, todavía se considera un desafío notable la aplicación de aprendizaje activo a tareas de reconocimiento de entidades y relaciones conjunto, sobre todo en escenarios con pocos recursos [160].

2.4.2. Entornos de Evaluación Competitivos

Una estrategia que se utiliza a menudo para fomentar la investigación sobre una tarea específica es la organización de una campaña de evaluación compartida. En contraste con la investigación regular, las campañas de evaluación a menudo tienen un marco de tiempo fijo y los recursos de evaluación no se revelan completamente (por ejemplo, las anotaciones para los conjuntos de prueba están ocultas) para permitir una comparación justa en un entorno competitivo amigable. En esta sección, analizamos algunos esfuerzos relevantes para la organización de campañas de evaluación tanto del dominio biomédico como para el manejo de la extracción de entidades y relaciones.

Varios servicios en línea permiten a los investigadores organizar desafíos y concursos de aprendizaje automático, proporcionando calificación automática, administración de usuarios y otras funciones útiles. Kaggle[4] es posiblemente la opción más popular, su principal limitación para nuestros propósitos es que para albergar un desafío, los organizadores deben comunicarse con los

[4] https://kaggle.com

proveedores de servicios. Las posibles alternativas son AIcrowd[5] y Codalab[6], que ofrecen opciones gratuitas para los organizadores del desafío.

El CLEF eHealth Evaluation Lab ha propuesto varios desafíos en el dominio biomédico, incluido el reconocimiento de entidades nombradas [18] y la extracción de información [19] en inglés, y ediciones posteriores en documentos franceses [20, 21]. En estos desafíos, los informes médicos de MEDLINE, EMEA y fuentes similares se anotan con trastornos, términos médicos, acrónimos y abreviaturas, que proporcionan escenarios de evaluación para varias tareas de PLN, incluido el reconocimiento de entidades, la normalización y la desambiguación. Otra tarea relevante la propone May and Priyadarshi [22] en Semeval 2017, centrada en el análisis sintáctico y la generación de AMR a partir de oraciones biomédicas en inglés. La aplicación de una conceptualización de propósito general, como AMR, a dominios específicos alentó a los participantes a cerrar la brecha entre el desarrollo de técnicas generalizables y la aplicación de heurísticas de dominio específico. Sin embargo, el análisis sintáctico de la estructura AMR es ya un problema complejo en sí mismo, que puede tener un impacto negativo en la participación de los investigadores en estos desafíos si no están especializados en este esquema. Los modelos más simples y de propósito general pueden fomentar un mayor grado de participación dada la curva de entrada más fácil. Un ejemplo de esto último es el Semeval 2017 Task 10 [23], un desafío en cuanto a la extracción de frases clave y relaciones de documentos científicos, con un modelo simple basado en tres clases de entidades y dos relaciones de propósito general. Esta tarea recibió una cantidad mucho mayor de presentaciones que la anterior, aunque ambos desafíos se realizaron en el mismo lugar y estaban dirigidos a audiencias similares.

Como se puede esperar, el inglés es el idioma más utilizado en los desafíos relacionados con NER, dado el mayor número de corpus y recursos disponibles. Sin embargo, se han realizado importantes esfuerzos para fomentar la investigación en idiomas menos destacados. Son relevantes para nuestro debate las campañas de IberLEF que se centran en los idiomas ibéricos, como el español, el portugués, el catalán y otras variaciones regionales. Dos ejemplos de tareas recientes relacionadas con NER son el desafío en para reconocimiento de entidades nombradas en portugués [161] y el desafío de

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

anonimización de documentos MEDDOCAN [162]. El primero propone el reconocimiento de entidades y la extracción de relaciones en el dominio general, en portugués. El segundo propone la identificación de menciones de entidades sensibles a la privacidad en documentos médicos, por ejemplo, nombres, direcciones, fechas, edades, etc.

Fuera del marco de una competencia, los sistemas de evaluación abiertos y de larga duración permiten a los investigadores evaluar sus enfoques con métricas de evaluación oficiales. Esto también puede proporcionar un repositorio centralizado del estado de la técnica, donde los enfoques existentes se resumen y se vinculan a los documentos existentes.

2.5. Discusión

El primer paso en la mayoría de las propuestas para el de descubrimiento de conocimiento consiste en la anotación de recursos lingüísticos. En la actualidad, existe un creciente interés por aplicar técnicas de aprendizaje no supervisado en gran escala con resultados sorprendentes en varias tareas de procesamiento de lenguaje natural [163]. Sin embargo, estos enfoques carecen aún de la capacidad para representar la semántica del texto subyacente de forma precisa [164]. Otra preocupación relevante en este contexto es la captura implícita de sesgos que se ve amplificada en tareas subsecuentes donde se emplean modelos de aprendizaje no supervisados entrenados en corpus masivos. Por tales motivo, en dominios de alto riesgo como la salud, donde capturar correctamente el significado semántico preciso de una frase puede significar la diferencia entre administrar un medicamento correcto o no, sigue siendo necesario definir representaciones explícitas del conocimiento.

La variedad de modelos de anotación y recursos lingüísticos en múltiples dominios emuestra que existe un balance entre la complejidad del modelo de anotación y su utilidad para ser usado en procesos automáticos de descubrimiento de conocimiento. Un modelo de anotación demasiado complejo, si bien es probable que represente con mayor precisión la semántica de un dominio, será más difícil de aprender automáticamente, y más costoso de producir suficientes recursos anotados para entrenar modelos de aprendizaje. Este es el caso de extremos como `AMR`, que están a un nivel semántico muy cercano al deseado para construir una ontología, pero son difíciles de generar

automáticamente. Por otro lado, un modelo demasiado simple, como tripletas sujeto-verbo-objeto, será incapaz de capturar las más importantes nociones semánticas del dominio de interés. Su principal desventaja es que todas las relaciones semánticas deberán ser desambigüadas en un paso posterior, pues la misma frase textual puede corresponder a diferentes conceptos en el dominio semántico. Por ejemplo, las relaciones ontológicas y teleológicas más comunes (hiponimia, holonimia, casualidad, implicación) tienen una infinidad de representaciones textuales posibles. El punto medio parece ser un modelo de anotación que define explícitamente relaciones de propósito general pero deje espacio para representar relaciones de dominio a partir de la superficie textual.

Capítulo 3

Resumen de los Resultados

En este capítulo se presenta un resumen de los principales resultados obtenidos durante esta investigación. Primero se presenta un esquema de anotación diseñado para capturar las entidades y relaciones semánticas fundamentales en un texto, independiente del idioma y dominio (Sección 3.1. Basado en este esquema de anotación se presentan cuatro versiones de un corpus anotado manualmente, en documentos fundamentalmente en idioma español del dominio de la salud, aunque también contienen una pequeña muestra de otros dominios e idiomas (Sección 3.2). A partir de estos recursos se diseña una tarea computacional que consiste en la detección automática de entidades y relaciones, con métricas y escenarios de evaluación que permiten una comparación objetiva entre diferentes sistemas (Sección 3.3). Con vistas a automatizar el proceso de evaluación de soluciones a esta tarea computacional, se propone una infraestructura que brinda implementaciones base (*baseline*) de algoritmos y todas las herramientas necesarias para desarrollar y evaluar implementaciones más complejas (Sección 3.4). Se presenta además la evaluación de un conjunto de sistemas propuestos por múltiples equipos de investigadores en cuatro ediciones del evento *eHealth Knowledge Discovery Challenge* (Sección 3.5), que han sido evaluados respectivamente en las cuatro versiones de los corpus anotados en esta investigación. Finalmente, a partir de la experiencia acumulada en la construcción de recursos lingüísticos, se diseña un entorno de anotación asistida basada en aprendizaje automático activo (Sección 3.6). A modo de resumen, la Figura 3.1 presenta

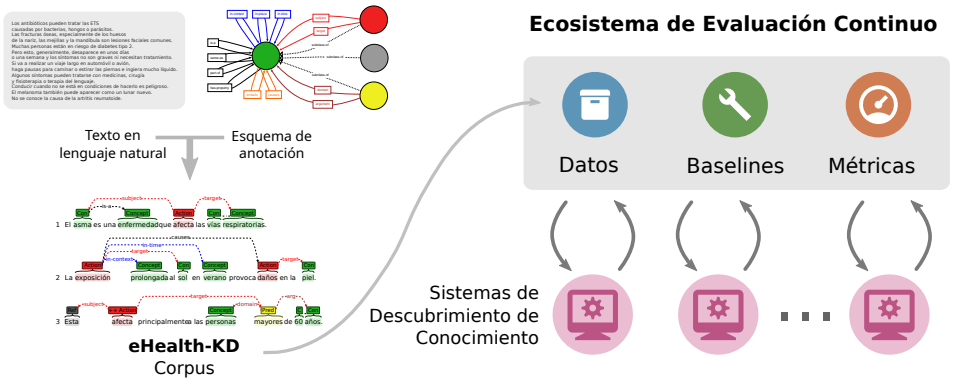


Figura 3.1: Esquema conceptual del ecosistema computacional diseñado en esta tesis.

un esquema conceptual del ecosistema computacional diseñado en esta Tesis.

3.1. Esquema de Anotación

El proceso de descubrimiento de conocimiento en lenguaje natural comienza por la identificación en el texto de los conceptos y relaciones más relevantes en un dominio. Para ello es necesario definir un modelo de representación semántico que capture estos conceptos. Dicho modelo se concreta en un esquema de anotación que permite a expertos humanos construir un corpus anotado semánticamente. Construir sistemas de descubrimiento de conocimiento automáticos generalizables a cualquier dominio requiere que los conceptos y relaciones representados sean de propósito general. Además, el esquema de anotación debe lograr un balance adecuado entre su capacidad expresiva y su simplicidad para ser anotado por expertos humanos y algoritmos de aprendizaje automático.

En esta Tesis se propone un esquema de anotación con estas características, que se inspira en varios recursos. En términos de representación del conocimiento, se basa en dos modelos diferentes para la conceptualización de la realidad: ontologías y teleologías. La Figura 3.2 muestra una anotación de ejemplo de tres oraciones con diferentes grados de complejidad. El esquema de anotación se explica en profundidad en el Capítulo 4.

La parte ontológica del esquema proporciona una representación de entidades en términos de sus relaciones jerárquicas y estructurales (es decir,

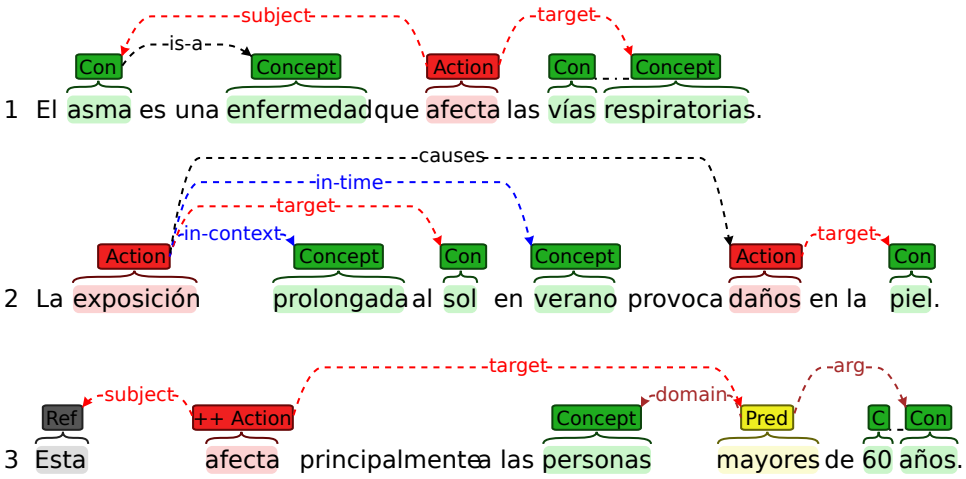


Figura 3.2: Ejemplo de anotación de tres oraciones. La anotación muestra las entidades y relaciones más relevantes definidas en el esquema de anotación propuesto.

`is-a`, `part-of`, `has-property` y `same-as`). Estas relaciones se basan en el diseño de ontologías de propósito general como ConceptNet [15] y YAGO [14]. La parte teleológica del esquema proporciona una representación de eventos o procesos que transforman entidades, es decir, representan las interacciones entre las entidades. Esta parte del esquema está respaldada en una estructura basada en el trabajo de Giunchiglia and Fumagalli [36]. El significado semántico exacto de estos conceptos y relaciones se explica con más detalle a continuación.

El elemento más importante de la anotación es la estructura Sujeto-Acción-Objetivo (*Subject-Action-Target*), que captura la interacción principal entre entidades en oraciones factuales. En esta interacción participan dos tipos de entidades fundamentales: `Concept` y `Action`. Un `Concept` define una entidad relevante en el dominio, que puede ser una sola palabra o múltiples tokens, contiguos o no. Una `Action` representa un proceso o evento causado por uno o más `Concept` (mediante la relación `subject`) y que impacta en uno o más `Concepts` (mediante la relación `target`). Las entidades conectadas a los roles `subject` y `target` también pueden ser de tipo `Action`, lo que permite que los conceptos simples se compongan en otros más complejos. La estructura Sujeto-Acción-Objetivo definida en este esquema se basa en una versión simplificada del marco teleológico propuesto por Giunchiglia and Fumagalli

[36]. Los elementos *Object* y *Action* en las teleologías están representados en este esquema por `Concept` y `Action` respectivamente. El rol *Function* en las teleologías, que expresa una instancia de un objeto que realiza una acción, puede equipararse aproximadamente al uso de una entidad de tipo `Action` ocupando el rol `subject` o `target` de otras acciones.

Una adición importante a este esquema de anotación es la entidad `Predicate`. Los predicados modelan la existencia de conceptos complejos (mediante el rol `domain`) que dependen de algunas condiciones previas (mediante el rol `arg`). Por ejemplo, en la Figura 3.2, Oración 3, el concepto de *personas mayores de 60 años* se puede definir con una anotación detallada, considerando "*personas*" como el dominio y "*60 años*" como argumento. Esta anotación permite la captura de información más detallada en lugar de simplemente anotar la frase completa como un concepto de varias palabras. Otra adición es `Reference`, que representa conceptos mencionados de forma implícita en una oración. Las palabras más comunes etiquetadas como referencias son: "*esto*", "*el*", "*la*", "*este*", es decir , generalmente pronombres y artículos.

Para refinar aún más la interpretación semántica de cada entidad, se define un conjunto de 4 atributos: `uncertain`, `emphasized`, `diminished` y `negated`. Estos atributos a menudo se pueden identificar mediante adjetivos u otros patrones sintácticos que aparecen en el contexto de una entidad determinada, pero en lugar de anotar toda la frase, la entidad correspondiente se etiqueta con el atributo. Por ejemplo, en la oración 3 de la Figura 3.2, la acción "*afecta*" se etiqueta con `emphasized`, debido a la presencia de la palabra "*principalmente*", y se representa en la anotación con un signo `++` en la entidad. El uso de atributos permite la captura de conceptos semánticos más refinados (es decir, grados de énfasis, negación, incertidumbre) mientras se mantiene el agnosticismo del idioma, ya que es irrelevante en qué parte del texto se presenta esa información. Puede aparecer explícitamente en una sola palabra (por ejemplo, un adjetivo) o implícitamente mediante frases idiomáticas u otros patrones lingüísticos sutiles. Estos atributos aumentan el rango semántico del esquema de anotación sin aumentar el número de tokens que deben ser anotados.

Este esquema define 4 relaciones ontológicas principales: `is-a`, `same-as`, `part-of` y `has-property`, con su semántica habitual, que pueden vincular cualquier par de entidades, simples o compuestas, entre sí. Estas relacio-

nes permiten la representación del conocimiento estructural, por ejemplo, conceptos relacionados en una estructura jerárquica y conceptos que son componentes de otros conceptos. Se define dos relaciones adicionales, `causes` y `entails`, para capturar causalidad e implicación lógica respectivamente. Estas relaciones, respectivamente de naturaleza teleológica y ontológica, son de gran importancia porque permiten la construcción de sistemas de razonamiento que pueden llegar a conclusiones y producir nuevos conocimientos a partir de un corpus existente.

Además, se definen 3 relaciones contextuales, para recopilar conocimientos importantes que generalmente aparecen como complementos gramaticales: `in-time`, `in-place`, `in-context`. La relación `in-time` se usa para expresar la duración de un evento. La relación `in-place` se usa para identificar una ubicación específica para una entidad de tipo `Action` o `Concept`. La relación `in-context` es una relación más genérica que representa una dependencia general entre dos entidades cuya naturaleza exacta no puede definirse por ninguna de las otras relaciones. Estas relaciones también son de naturaleza teleológica, ya que no definen una afirmación en sí, sino que son útiles para especificar las condiciones en las que ocurren algunos eventos. Por ejemplo, en la Oración 2 (Figura 3.2, la anotación *exposición* ⇒ `in-context` ⇒ *prolongado* no implica que el concepto "*exposición*" incondicionalmente tenga la calidad "*prolongado*". Solo cuando este concepto complejo se usa como `subject` o `target` de una entidad de tipo `Action` o en otra relación, es que la contextualización adquiere un significado.

La figura 3.3 resume el esquema de anotación. Este esquema está diseñado para capturar el conocimiento semántico más relevante presente en un corpus arbitrario. Por esta razón, no se definieron relaciones o entidades específicas de dominio (es decir, no hay entidades específicas para enfermedades, pacientes, tratamiento, etc.). Por el contrario, las relaciones específicas de dominio pueden representarse mediante acciones y sus roles correspondientes.

3.2. Corpus

Basado en el esquema de anotación definido en la Sección 3.1 se anotaron un total de 4 versiones incrementales de un corpus en idioma español. Cada versión corresponde a una edición de la campaña de anotación *eHealth-KD*

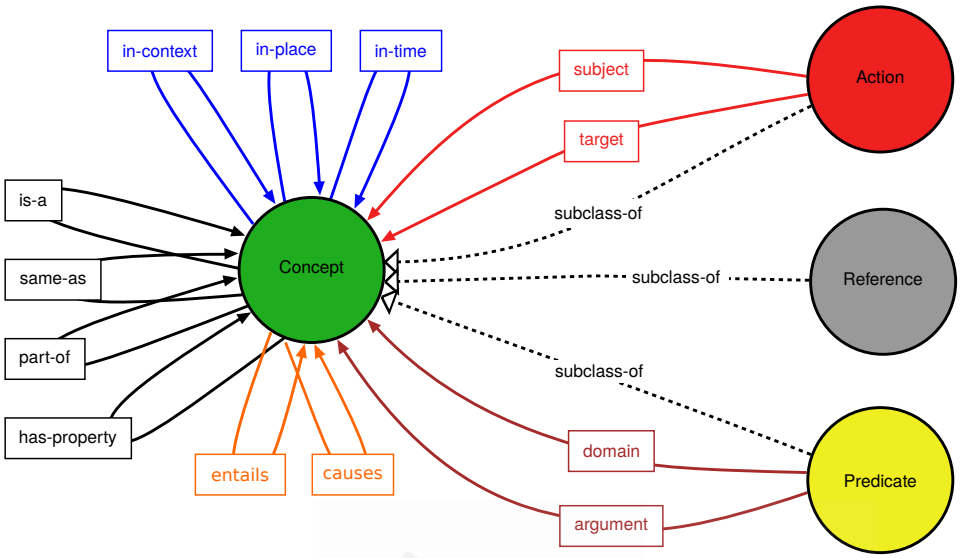


Figura 3.3: Diagrama resumen del esquema de anotación general.

que se presenta en la Sección 3.5. La primera versión se anotó con una variante ligeramente diferente del esquema de anotación, debido a que este esquema fue evolucionando durante el proceso de investigación. En esta Sección se presentan las características fundamentales de este corpus, incluyendo la estrategia de anotación seguida en cada caso.

La Figura 3.4 describe el proceso de anotación seguido en todos los corpus desarrollados en esta investigación, consistente en 4 etapas fundamentales. En primer lugar un conjunto reducido de anotadores expertos construye una colección de prueba con un número pequeño de oraciones (p.e., 50 oraciones), escogidas específicamente para cubrir la mayor cantidad posible de patrones de anotación. Con esta colección se construye una guía de anotación que se distribuye entre un conjunto de anotadores no-expertos para su entrenamiento y evaluación. Los anotadores no-expertos realizan la anotación manual del resto del corpus, de forma que cada oración es anotada por exactamente dos personas diferentes sin interacción entre sí. Luego se realiza un proceso de mezcla para escoger entre las 2 versiones anotadas de cada oración. Para cada oración anotada, esta mezcla la realiza respectivamente uno de los anotadores no-expertos que no haya participado en la primera fase en dicha oración. De esta forma cada oración ha sido evaluada hasta este paso por 3 personas diferentes. Finalmente, los anotadores expertos revisan en conjunto todas las oraciones anotadas y proponen modificaciones menores que deben

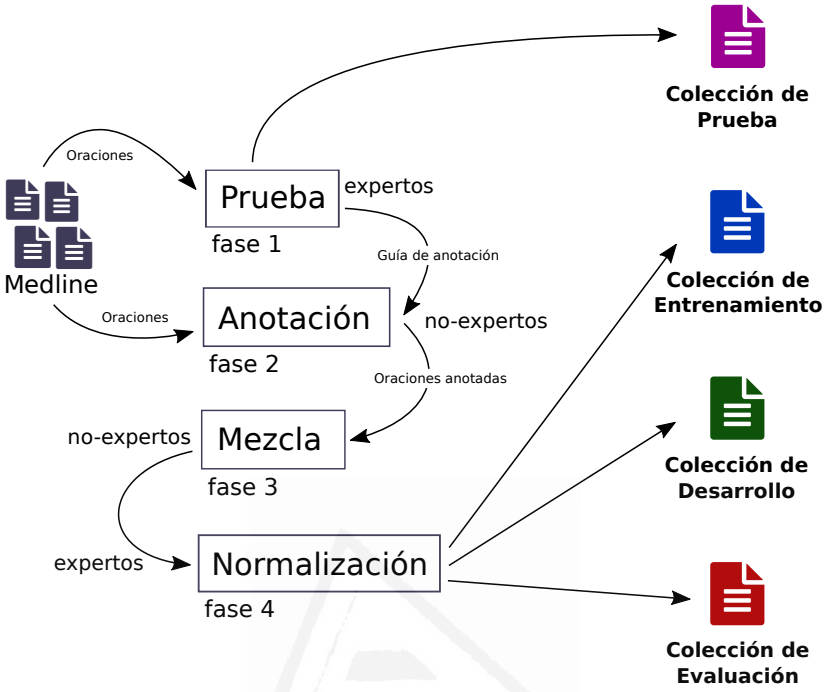


Figura 3.4: Esquema general del proceso de anotación.

ser aprobadas por unanimidad.

Siguiendo este proceso de anotación y el esquema definido en la Sección 3.1, se han construido 4 ediciones del corpus, de tamaño similar, utilizados respectivamente como escenarios de evaluación en las ediciones 2018, 2019, 2020 y 2021 del *eHealth-KD Challenge*. Aunque la composición exacta del comité de anotadores no-expertos ha variado en cada una de las ediciones, el conjunto de anotadores expertos se ha mantenido constante, y la mayoría de los anotadores no-expertos han participado en más de una edición. En las cuatro ediciones el grueso de las oraciones se han obtenido de la base de datos *Medline Plus*[1] que contiene artículos en idioma español (además de inglés) en el dominio de la salud. En la edición 2020 se adicionó como fuente de información un conjunto de 200 artículos seleccionados aleatoriamente de *Wikinews* para estimular el desarrollo de sistemas de aprendizaje multi-dominio y evaluar cuán generalizable es el esquema de anotación. En la edición 2021, con motivo de la agravada situación epidemiológica a nivel global

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

propiciada por la pandemia del COVID-19, se adicionaron 200 oraciones del corpus `CORD-19` [165] relacionadas con esta enfermedad, en idioma inglés. De esta forma, la mayoría de las oraciones anotadas son en idioma español, excepto la última colección proveniente de `CORD-19`.

La Tabla 3.1 resume las estadísticas fundamentales de los tres corpus anotados, denominados respectivamente *eHealth-KD* 2018, 2019, 2020 y 2021. El corpus de la edición 2018 (ver Capítulo 5) fue anotado con una versión simplifica del esquema de anotación, dado que aún no se había formalizado completamente. Por este motivo solamente se anotaron acciones, conceptos, y 4 relaciones semánticas. En total se anotaron 1,173 oraciones que agrupan 13,113 elementos semánticos. En la edición 2019 (ver Capítulo 7) se introdujeron todos los elementos semánticos del esquema de anotación y se anotaron 1,045 oraciones nuevas con un total de 13,246 elementos. Para la edición 2020 se reutilizaron 1,000 oraciones de la edición anterior, y se adicionaron 500 oraciones nuevas siguiendo el mismo esquema de anotación, 300 de dominio médico y 200 de artículos periodísticos. Para la edición 2021 se reutilizaron 1650 oraciones de la versión anterior, y se anotaron 350 oraciones nuevas, de ellas 200 en idioma inglés. En total se han anotado 3,068 oraciones en idioma español con 41,163 elementos semánticos que capturan múltiples patrones lingüísticos en documentos del dominio médico y periodístico, en idiomas español e inglés.

3.3. Definición de Tareas

Los recursos lingüísticos presentados en la Sección 3.2 son necesarios para entrenar sistemas computacionales que sean capaces de extraer automáticamente contenido semántico de lenguaje natural. Para avanzar en esta tarea es conveniente diseñar un entorno de evaluación que permita comparar de forma objetiva los enfoques posibles. Para evaluar mejor las fortalezas y debilidades de los diferentes enfoques, la tarea de anotación automática se divide en dos subtareas:

Subtarea A: Reconocimiento de Entidades. El propósito de esta subtarea es identificar todas las entidades mencionadas en una oración y sus clases correspondientes. (i.e., `Concept`, `Action`, `Predicate` y `Reference`.

	Corpus eHealth-KD				
	2018	**2019**	**2020**	**2021**	**Total**
Oraciones	1,173	1,045	*1,500	*1,900	3,068
Prueba	29	45	0	0	74
Entrenamiento	559	600	*800	*1500	1,159
Desarrollo	285	100	*200	*100	535
Evaluación	300	300	500	*300	1,400
Anotaciones	13,113	13,246	8,538	6,266	41,163
Entidades	7,188	6,612	4,237	3,162	21,199
`Concept`	5,366	4,092	2,803	2,366	14,627
`Action`	1,822	1,742	944	548	5,056
`Predicate`	-	563	420	225	1,208
`Reference`	-	215	70	23	308
Relaciones	5,925	6,049	4,037	2,934	18,945
`target`	2,120	1,729	855	483	5,187
`subject`	1,466	894	706	340	3,406
`is-a`	1,057	566	383	241	2,247
`part-of`	393	94	57	28	572
`has-property`	836	159	111	272	1,378
`same-as`	53	124	85	43	305
`in-context`	-	677	584	674	1,935
`in-place`	-	400	351	285	1,036
`in-time`	-	165	215	120	500
`domain`	-	364	285	141	790
`argument`	-	343	233	111	687
`causes`	-	367	131	103	601
`entails`	-	167	41	72	280
Attributos	-	585	264	170	1,019
`diminished`	-	18	11	13	42
`emphasized`	-	124	80	59	263
`negated`	-	164	72	41	277
`uncertain`	-	279	101	57	437

Tabla 3.1: Estadísticas de las 4 ediciones del corpus *eHealth-KD*. Los números anotados con * en 2020 y 2021 indican que se han reutilizado oraciones de las ediciones anteriores. En este caso solo se contabilizan en el total las anotaciones las oraciones nuevas anotadas en cada edición.

Subtarea B: Extracción de relaciones. El propósito de esta subtarea es detectar todas las relaciones semánticas entre cada par de entidades ya etiquetadas en cada oración.

Como criterio de evaluación de ambas subtareas se propone una versión extendida de la métrica F_1 modificada para tratar coincidencias parciales. La métrica F_1 depende de las anotaciones correctas, incorrectas, parciales, faltantes y espurias en todo el conjunto de evaluación. Dependiendo de la(s) subtarea(s) bajo evaluación, definimos los siguientes tipos de resultados:

Subtarea A - Correctos C_A: cuando una anotación coincide exactamente con la anotación correcta correspondiente.

Subtarea A - Incorrectos I_A: cuando una anotación coincide con una anotación correcta con respecto al espacio de texto pero define una etiqueta de entidad diferente.

Subtarea A - Parciales P_A: cuando un fragmento de texto tiene una intersección no vacía pero inexacta con una anotación correcta, como el caso de "*vías respiratorias*" y "*vías*" en la Figura 3.2, oración 2. Las frases parciales solo se comparan con una frase correcta (es decir, la primera frase parcialmente coincidente desde el principio de la oración) para evitar que algunos fragmentos de texto grandes que cubren la mayor parte del documento obtengan una puntuación muy alta.

Subtarea A - Faltantes M_A: cuando no se produce una anotación que aparece en la colección de anotaciones correctas.

Subtarea A - Espurios S_A: cuando se produce una anotación que no aparece en la colección de anotaciones correctas.

Subtarea B - Correctos C_B: cuando existe una relación entre dos entidades en la colección de anotaciones correctas.

Subtarea B - Faltantes M_B: cuando no se produce una relación en la colección de anotaciones correctas.

Subtarea B - Espurios S_B: cuando se produce una relación pero no aparece en la colección de anotaciones correctas.

Se define *Precision*, *Recall*, y F_1 de la manera usual, teniendo en cuenta que para cada escenario de evaluación solo se consideran los términos relacionados con la(s) subtarea(s) bajo evaluación.

$$Precision = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + S_A + S_B} \tag{3.1}$$

$$Recall = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + M_A + M_B} \tag{3.2}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.3}$$

Finalmente, se propone F_1 como se define en la Ecuación 3.3 como la métrica oficial para comparar diferentes enfoques. En cada versión del corpus *eHealth-KD* se define un subconjunto diferente de oraciones para evaluar cada subtarea de forma independiente o ambas tareas de manera simultánea.

3.4. Infraestructura de Aprendizaje y Evaluación

Para apoyar a los investigadores en el desarrollo de tecnologías de descubrimiento de conocimiento, se proporciona un conjunto de herramientas e infraestructura que permiten un proceso de experimentación más rápido y objetivo. Estos recursos están disponibles gratuitamente para la comunidad científica en una colección de repositorios de Github[2]. El conjunto de recursos desarrollado consiste en los siguientes elementos:

- Archivos de texto plano y anotaciones en formato BRAT Standoff [101] para el corpus *eHealth-KD*, dividido en colecciones de entrenamiento, desarrollo y evaluación.
- Archivos de configuración necesarios para desplegar un servidor BRAT con el objetivo de analizar y ampliar el corpus *eHealth-KD*, o para crear otros recursos lingüísticos basados en el esquema de anotación descrito en la Sección 3.1.

[2] https://ehealthkd.github.io

- Herramientas en el lenguaje de programación Python para cargar y manipular las anotaciones de BRAT Standoff así como para producir resultados formateados correctamente.
- Herramientas para configurar y ejecutar algoritmos de evaluación para la tarea definida en la Sección 3.3, incluidas las subtareas, y calcular las métricas de evaluación oficiales.
- Un conjunto de implementaciones de algoritmos de aprendizaje básicos con diferentes grados de complejidad, incluida una estrategia aleatoria y varios enfoques clásicos.

Usando las herramientas antes mencionadas, los investigadores pueden desarrollar rápidamente nuevos enfoques extendiendo las implementaciones *baseline* o desarrollando una solución desde cero, sin tener que lidiar con la configuración del entorno o la implementación de las métricas de evaluación. Además de poder evaluar sus soluciones personalmente, los investigadores también pueden subir sus soluciones a un entorno de evaluación en la nube y obtener automáticamente las métricas relevantes así como comparar sus resultados con soluciones ya publicadas. Se mantiene una tabla de clasificación oficial que sirve como un estado del arte actualizado en todas las subtareas.

3.5. Evaluación de Sistemas

Las cuatro versiones del corpus *eHealth-KD* presentadas en la Sección 3.2 han sido utilizadas en la evaluación de sistemas de extracción de conocimiento diseñados por diversos equipos de investigadores en el marco del evento *eHealth Knowledge Discovery*, organizado en los años 2018, 2019, 2020 y 2021. En la Tabla 3.2 se resumen los resultados más importantes de las cuatro ediciones de este evento. Se reportan el número de participantes (equipos de investigadores), los mejores resultados alcanzados tanto en el proceso completo de extracción como en cada una de las subtareas definidas en la Sección 3.3 (en términos de la métrica F_1), así como un resumen de las técnicas más utilizadas.

Con respecto a los resultados obtenidos en la solución de las tareas, es necesario reconocer que la complejidad de la edición 2018 es significativamente menor debido a que existe un menor número de tipos de entidades y relaciones.

	2018	**2019**	**2020**	**2021**
Equipos Registrados	31	30	26	30
Equipos Participantes	6	10	8	9
Resultados generales	0.744	0.639	0.660	0.531
Subtarea A (entidades)	0.872	0.820	0.825	0.706
Subtarea B (relaciones)	0.448	0.626	0.633	0.430
Técnicas				
PLN clásico	5	7	0	3
Reglas	2	2	3	1
Embeddings	2	7	8	9
Redes neuronales	3	10	7	9
Transformers	0	1	5	8
Conocimiento externo	2	1	3	9
Solución *end-to-end*	0	1	2	3

Tabla 3.2: Resumen de los resultados de las tres ediciones del *eHealth Knowledge Discovery* y las técnicas más comúnmente aplicadas en cada edición.

Por este motivo los resultados no son directamente comparables con los de las ediciones 2019, 2020 y 2021. De la misma forma, la edición 2021 incluyó tres dominios y dos idiomas, por lo que es de esperar que los resultados sean ligeramente inferiores a las ediciones anteriores. Aún así, una posible conclusión de estos resultados es que la detección de entidades está de manera general resuelta mientras que la extracción de relaciones consiste en el mayor reto.

Analizando las técnicas más utilizadas por los sistemas presentados en cada edición, se puede notar una tendencia hacia los sistemas *end-to-end* basados en arquitecturas de redes neuronales, o sea, sistemas que resuelven ambas tareas como parte de una misma arquitectura en vez de como pasos separados. De manera general las redes neuronales, y en particular los *embeddings* (e.g., word2vec, glove) y las arquitecturas *transformer* (e.g., BERT, GPT), han reemplazado a las técnicas clásicas de procesamiento de lenguaje natural. Sin embargo, aún es necesario el uso de ciertas reglas específicas del dominio, por ejemplo, para resolver el problema de las entidades superpuestas.

Por otro lado, los enfoques que aprovechan conocimiento externo, aunque no son mayoría siguen teniendo importancia, fundamentalmente en la forma de *embeddings* de dominio específico.

3.6. Anotación Asistida

El proceso de anotación, mezcla y normalización que se ha llevado a cabo durante el transcurso de esta investigación ha permitido reconocer la importancia de esta fase como parte del desarrollo de sistemas de aprendizaje automático y descubrimiento de conocimiento. La construcción de recursos lingüísticos anotados por expertos es fundamental para entrenar este tipo de sistemas, y consiste en una de las tareas más costosas en términos de esfuerzo humano. Además, si no se cuenta con una planificación ideal, es probable que los recursos creados sean menos efectivos. Por ejemplo, pueden contener oraciones repetidas o semánticamente similares, conjuntos desbalanceados de etiquetas, y contradicciones internas. Una forma de aliviar estos problemas y acelerar el proceso de anotación es introducir un mecanismo de aprendizaje automático dentro del propio ciclo de anotación, mezcla y normalización, que ayude a seleccionar los mejores ejemplos a anotar y garantice que no existe redundancia innecesaria en el corpus.

Como parte de esta investigación se propone una estrategia de aprendizaje activo diseñada para un corpus arbitrario de oraciones en lenguaje natural, cada una de las cuales debe ser anotada por un experto humano a nivel semántico, tal y como ocurre con el esquema de anotación propuesto en la Sección 3.1. Con el objetivo de diseñar una propuesta generalizable, se considera un conjunto arbitrario $\mathcal{E}$ de etiquetas de entidades, cada una de las cuales puede abarcar uno o más tokens, continuos o discontinuos, y un conjunto predefinido $\mathcal{R}$ de tipos de relaciones binarias entre entidades. Este modelo de anotación abstracto puede representar una amplia gama de tareas diferentes, incluyendo el corpus *eHealth-KD* en el cuál se inspira, pero abarcando además desde la extracción de relaciones de dominio específico (por ejemplo, interacción gen-proteína) hasta la representación semántica de propósito general (por ejemplo, análisis de AMR). Esta estrategia se describe en detalle en el Capítulo 9.

La estrategia de aprendizaje activo propuesta en esta investigación fun-

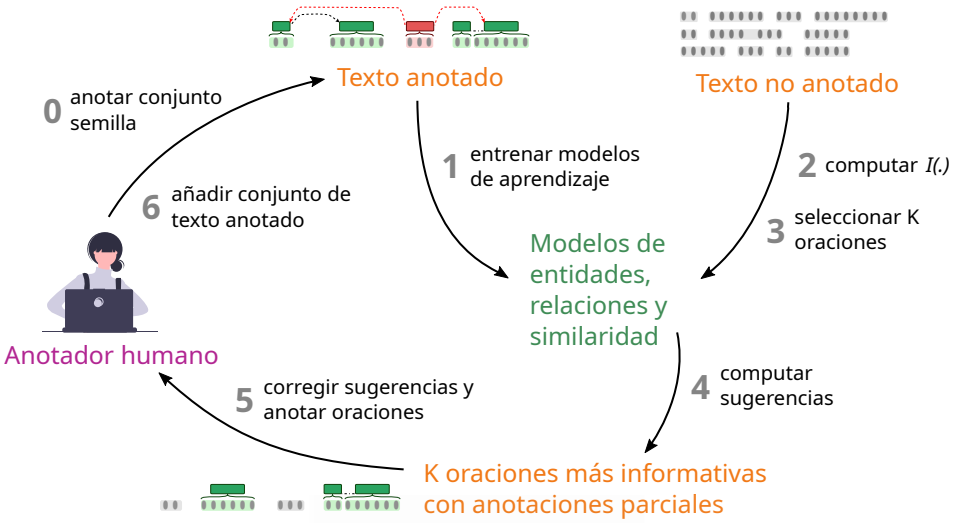


Figura 3.5: Ilustración de alto nivel de la estrategia de aprendizaje activo presentada en esta investigación para la anotación asistida del corpus ehealth-KD.

ciona de forma iterativa en lotes de K oraciones (por ejemplo, $K = 10$). La figura 3.5 muestra un resumen del proceso. En cada iteración, existe un conjunto etiquetado $\mathbf{L}$ con $|\mathbf{L}| = n \times K$ oraciones anotadas manualmente por un anotador humano, y un conjunto mayor $\mathbf{U}$ de oraciones no etiquetadas. Inicialmente, el anotador humano selecciona K oraciones representativas y realiza una anotación manual completa (paso 0). Posteriormente, dos modelos de aprendizaje automático se entrenan iterativamente en las oraciones etiquetadas manualmente (paso 1) y una métrica de *informatividad*, $I(s)$, se calcula para cada oración $s \in \mathbf{U}$ (paso 2). Las mejores oraciones K en términos de $I(\cdot)$ se seleccionan (paso 3) y el modelo produce una predicción de entidades y relaciones para cada una (paso 4). Cada predicción tiene una métrica de *incertidumbre* asociada, $H(\cdot)$, estimada por los modelos. En función de esta incertidumbre y umbrales predefinidos u_e y u_r para entidades y relaciones respectivamente, todas las entidades e_i (relaciones r_j) con una incertidumbre estimada $H(e_i) > u_e$ ($H(r_j) > u_r$) se descartan. Finalmente, las oraciones seleccionadas y parcialmente anotadas se presentan al anotador humano, quien debe corregir las anotaciones incorrectas y agregar las anotaciones faltantes (paso 5). Las oraciones corregidas se incorporan al grupo etiquetado para la siguiente iteración (paso 6).

Para estimar la informatividad $I(s_i)$ de cada oración $s_i \in \mathbf{U}$, se define una métrica basada en el modelo de *uncertainty sampling* [153]. Dado un conjunto de n anotaciones de entidades $E_i \subseteq \mathcal{E}^n = \{e_1^i, \ldots, e_n^i\}$ y m anotaciones de relaciones $R_i \subseteq \mathcal{R}^m = \{r_1^i, \ldots, r_m^i\}$ producidas para una oración s_i, se define la incertidumbre de cada entidad e_k^i (o relación r_k^i) como la entropía de la distribución de probabilidades de cada una de las posibles etiquetas para dicha entidad o relación. Formalmente:

$$H(e_k^i) = -\sum_{l_j \in \mathcal{E}} P(e_k^i = l_j | s_i; \theta) \log_2 P(e_k^i = l_j | s_i; \theta)$$
$$H(r_k^i) = -\sum_{l_j \in \mathcal{R}} P(r_k^i = l_j | s_i; \theta) \log_2 P(r_k^i = l_j | s_i; \theta)$$

Donde θ representa los parámetros del modelo de aprendizaje que se utiliza para estimar estas probabilidades.

De la misma forma se puede definir la incertidumbre media asociada a las entidades y relaciones producidas, respectivamente:

$$\hat{H}(E_i) = \frac{1}{n} \sum_{e_k^i \in E_i} H(e_k^i) \qquad \hat{H}(R_i) = \frac{1}{m} \sum_{r_k^i \in R_i} H(r_k^i)$$

Además, se define una métrica de densidad de información $ID(s_i)$ para estimar cuán representativa es cada oración s_i con respecto al conjunto de oraciones anotadas. De forma similar a Settles and Craven [153], $ID(s_i)$ se define como la similaridad promedio de la oración s_i a un conjunto K de oraciones anotadas:

$$ID(s_i) = \frac{1}{K} \sum_{s_j \in \mathbf{L}_i^*} sim(s_i, s_j)$$

Donde $\mathbf{L}_i^*$ es el subconjunto de las K oraciones anotadas que maximizan la similaridad con respecto a s_i. Cualquier métrica de similaridad podría ser utilizada, pero en esta investigación se propone usar *embeddings* de *Doc2Vec* [166] pre-entrenados en el conjunto $\mathbf{U}$ de oraciones no anotadas.

Finalmente, la *informatividad* total de una oración no anotada s_i se estima a partir de la incertidumbre de cada componente (entidad o relación) ponderados por la densidad de información de la oración:

$$I(s_i) = \left[\hat{H}(E_i) + \hat{H}(R_i)\right] \times ID(s_i)^\beta$$

Para evaluar la efectividad de este enfoque se simula el proceso de anotación asistida en el corpus *eHealth-KD* 2019, comparando las estrategias de aprendizaje activo con la anotación en el orden original sin sugerencias (como modelo de base). Como el proceso de anotar un corpus es costoso, la simulación se basa en las anotaciones correctas de la colección de entrenamiento. La mejora se puede estimar comparando cuántas oraciones necesitan anotaciones para alcanzar un rendimiento específico de los algoritmos de aprendizaje automático (medido en términos de F_1 en la colección de pruebas).

Para ilustrar el grado de reducción de tiempo alcanzado, la Figura 3.6 muestra el número mínimo de oraciones que deben anotarse para alcanzar diferentes puntajes relativos de F_1. Por ejemplo, después de anotar las primeras 400 oraciones, es posible lograr un 95 % del resultado final de F_1 cuando se utiliza todo el corpus. Sin embargo, para alcanzar la puntuación objetivo, las primeras 880 oraciones de las 1000 totales deben anotarse si el corpus se anota en el orden original (seguido por el modelo de base). Por el contrario, la estrategia de aprendizaje activo solo requiere anotar entre 530 y 580 oraciones para alcanzar el mismo valor de F_1, ahorrando así entre 35 % y 40 % del tiempo de anotación humano.

Otro análisis interesante es estimar hasta qué punto las anotaciones sugeridas reducen aún más el tiempo total de anotación. Un anotador humano que use esta herramienta deberá aceptar algunas de las anotaciones sugeridas, corregir las incorrectas y anotar las faltantes. Cada una de estas acciones tiene un costo diferente. Para cuantificar la mejora en el tiempo general que producen las sugerencias, se asigna un costo relativo (en términos de unidades de tiempo abstractas) a cada uno de los siguientes tipos de anotaciones:

Faltantes: 1 unidad de tiempo.

Espurias: 2 unidades de tiempo.

Correctas: 0.25 unidades de tiempo.

Parciales: 0.5 unidades de tiempo.

Esta estructura de costos asume que el problema de corregir anotaciones incorrectas es más complejo que simplemente producir las anotaciones correctas, al tiempo que reconoce que incluso estar de acuerdo con las anotaciones correctas tiene un costo. Para que una estrategia de aprendizaje activo sea

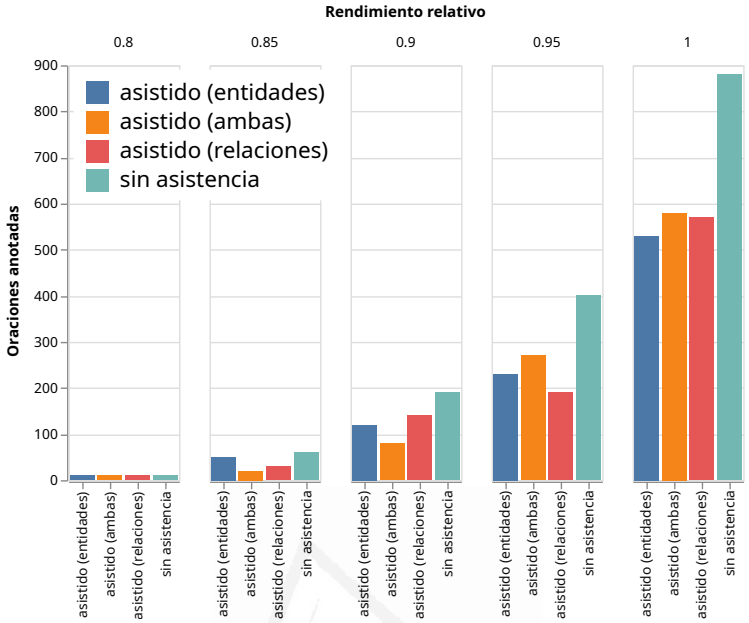


Figura 3.6: Número mínimo de oraciones necesarias para alcanzar un rendimiento específico en relación con la métrica F_1 con cada estrategia de aprendizaje activa para la sugerencia de oraciones.

útil, debe proporcionar suficientes anotaciones correctas para superar el costo de corregir las anotaciones incorrectas; por lo tanto, debe priorizar la precisión sobre el recobrado. Una simulación de este proceso para diferentes valores de los umbrales de incertidumbre arroja que, en el caso óptimo, se reduce a un 76 % el costo de arreglar una oración parcialmente anotada con respecto a la anotación manual.

Se puede estimar una reducción general en el tiempo de anotación para esta simulación experimental combinando las mejoras proporcionadas por las sugerencias de anotación y el orden de las oraciones. Asumiendo que ambos efectos son independientes, el mejor de los casos para este corpus sugiere lo siguiente. Usando el enfoque de aprendizaje activo, un anotador humano habría necesitado anotar solo 530 oraciones de 1000, cada una de ellas con un costo de tiempo estimado de 76 % en comparación con la anotación completa. Esto da como resultado una reducción general de hasta 60 % del tiempo de anotación total, produciendo un corpus más pequeño en el que los modelos de aprendizaje automático aún pueden ser entrenados, ofreciendo el mismo

rendimiento que los modelos entrenados en el corpus original.

3.7. Discusión

Esta sección presenta una discusión general de las principales contribuciones de esta investigación, las lecciones aprendidas y las limitaciones de las soluciones actuales propuestas para la tarea *eHealth Knowledge Discovery*. También se destacan ideas interesantes para la investigación futura, basadas en los conocimientos obtenidos al analizar los enfoques más prometedores.

3.7.1. Contribuciones

Se han desarrollado diferentes representaciones semánticas para capturar el conocimiento expresado en lenguaje natural (p.e., `AMR`, `FrameNet` y `PropBank`, ver Sección 2.2). El principal inconveniente de estas representaciones es su complejidad, ya que a menudo dependen de lexicones que definen los roles semánticos específicos para cada palabra. Por lo tanto, desarrollar sistemas de inteligencia artificial para el descubrimiento del conocimiento con este nivel de detalle es aún un problema no resuelto. El uso de representaciones semánticas más simples, que no se basen en roles o relaciones específicas, puede simplificar el desarrollo de técnicas basadas en el aprendizaje automático.

Esta investigación propone una línea de desarrollo en esta dirección, donde el descubrimiento de conocimiento con un alto nivel de abstracción pueda ser refinado posteriormente para tareas de dominio específico. El propósito no es reemplazar las representaciones semánticas detalladas, como `AMR` o `FrameNet`, sino proporcionar una representación más general que se pueda usar como un paso inicial en diferentes tareas de descubrimiento de conocimiento. Este tipo de representación semántica puede simplificar tareas posteriores como el aprendizaje de ontologías, de la misma manera que el etiquetado POS-tag de propósito general a menudo se realiza antes de tareas de PLN más complejas. Basado en las experiencias obtenidas durante el proceso de anotación, mezcla y normalización, se ha diseñado un mecanismo para la anotación semi-automática. En los experimentos realizados se pudo comprobar cómo utilizando modelos de aprendizaje relativamente sencillos se puede acelerar considerable el proceso de anotación.

Los recursos, herramientas e infraestructura desarrollados en esta investigación tienen como objetivo proporcionar una base para que la comunidad científica construya técnicas de descubrimiento de conocimiento de propósito general. Progresar en esta dirección depende no solo de los avances teóricos, como mejores arquitecturas de aprendizaje profundo o técnicas de procesamiento del lenguaje natural, sino también de la disponibilidad de recursos que permitan una experimentación eficiente. En este sentido, esta propuesta introduce una nueva tarea de descubrimiento de conocimiento así como métricas de evaluación formalmente definidas y un conjunto de evaluación práctico donde los investigadores pueden desarrollar rápidamente nuevas técnicas y obtener retroalimentación inmediata. También es un paso en la dirección de alentar la investigación de descubrimiento de conocimiento en idiomas menos utilizados, como el español, y en dominios socialmente importantes como la salud, así como evaluar las capacidades de generalización de los sistemas existentes a múltiples dominios.

3.7.2. Desafíos actuales y futuros

Los resultados de las cuatro ediciones del evento *eHealth-KD* permiten analizar la complejidad de los diversos pasos involucrados en el diseño de sistemas de descubrimiento automático de conocimiento para esta tarea. La mayoría de los sistemas modelaron la tarea como una secuencia en la que se reconocen primero las entidades y luego se extraen las relaciones. La subtarea A se modela comúnmente como un problema de etiquetado de secuencia y se resuelve mediante técnicas estándar, por ejemplo, redes Bi-LSTM y CRF. La subtarea B se modela comúnmente como un problema de clasificación estándar, donde la entrada consiste en una representación del un par de conceptos, utilizando *embeddings* contextuales y otras características sintácticas. Sin embargo, algunos de los sistemas de mejor desempeño en las últimas ediciones del *eHealth Knowledge Discovery* consisten en enfoques *end-to-end* que predicen las entidades y relaciones simultáneamente. Además de diferencias marginales en las arquitecturas y las metodologías de entrenamiento, argumentamos que la fortaleza de estos sistemas surge del efecto de regularización del aprendizaje de una representación unificada para ambas subtareas, en lugar de diferentes representaciones, lo que permite obtener más información de la misma cantidad de datos de entrenamiento.

En base a estas observaciones, estimamos que los enfoques más efectivos para este problema deberían considerar las siguientes estrategias: resolver ambos problemas simultáneamente en lugar de secuencialmente; usando *embeddings* pre-entrenados de propósito general o contextuales en lugar de *embeddings* personalizados; aplicar técnicas de ampliación del conjunto de datos para incrementar la cobertura estadística; y, diseñar reglas específicas para lidiar con la superposición y la discontinuidad de entidades.

En comparación con el nivel de experticia humano, la subtarea B parece ser considerablemente más difícil para los sistemas de aprendizaje automático que para los humanos. En experimentos reportados en los Capítulos 7 y 8, un anotador experto supera al sistema con el mejor rendimiento en un rango entre 8,8 % y 12,5 % en la tarea completa, pero solo en un 4,1 % en la subtarea A, en comparación con hasta un 19,1 % en la subtarea B. Intuitivamente, la subtarea B debería ser más difícil, ya que el número de etiquetas para predecir es mayor que en la subtarea A. Sin embargo, esto no explica la diferencia en el rendimiento entre humanos y sistemas de aprendizaje automático. En promedio, los sistemas que intentan resolver la subtarea B obtienen un valor F_1 a menudo significativamente menor en la subtarea B en comparación con la tarea completa, mientras que el anotador experto es similar o ligeramente mejor en la subtarea B. Esto indica que los humanos pueden obtener conocimiento adicional al ver las anotaciones correctas para la subtarea A que los sistemas de aprendizaje automático no reconocen. Sin embargo, el hecho de que la subtarea B sea significativamente más difícil para los humanos que la subtarea A es una indicación del alto grado de análisis cualitativo involucrado en este problema. Como tal, hay un umbral por encima del cual incluso los expertos humanos no estarán completamente de acuerdo, dada la naturaleza inherentemente subjetiva de la comprensión del lenguaje natural.

La aparición de las arquitecturas *Transformer* y su éxito reciente en varias tareas de PLN [167] abre las puertas a mejorar potencialmente los resultados actuales con poco esfuerzo adicional. La primera edición del evento (en 2018) consistió principalmente en sistemas híbridos, usando una combinación de técnicas de PLN basadas en reglas y conocimiento externo con aprendizaje automático. Sin embargo, la edición de 2019 no incluyó casi ningún enfoque basado en reglas, en favor de arquitecturas de aprendizaje profundo más complejas, mientras las ediciones de 2020 y 2021 se destacaron por el uso

casi exclusivo de *Transformers* combinados con diseños de arquitecturas específicas para cada subtarea.

Sin embargo, se aprecia que todavía hay un gran margen de mejora a través de enfoques que consideren la información global de la oración completa en lugar de simplificar el problema como un conjunto de subtareas de clasificación independientes. Desde una perspectiva humana, la anotación de una oración es un proceso global, en el que la decisión de considerar una palabra específica como `Action` o `Predicate` hace que un anotador reconsidere la oración completa y potencialmente cambie otras anotaciones. La incorporación de este tipo de conciencia global en un sistema requiere más que *embeddings* contextuales o incluso modelos de lenguaje a nivel de oración. El sistema debe poder evaluar una oración anotada de manera incompleta y potencialmente deshacer o corregir etiquetas anteriores a medida que avanza, hasta que se alcance un criterio de convergencia adecuado. Este tipo de comportamiento requiere un marco más expresivo que el que ofrecen las arquitecturas de aprendizaje supervisado puro. Un enfoque posible consiste en diseñar un agente anotador que observe la oración completa y realice acciones similares a cómo los humanos abordan este problema, posiblemente a través de aprendizaje por refuerzo.

Otra consideración importante es la alta correlación entre la identificación correcta de cada tipo de entidad y relación con su frecuencia relativa en el conjunto de entrenamiento. Esto refuerza la idea de que la mayoría de los enfoques actuales básicamente realizan un aprendizaje estadístico puro y, por lo tanto, no son capaces de capturar con precisión los matices semánticos de cada una de estas etiquetas. Esta evidencia también apunta a la necesidad de enfoques más conceptuales que realmente intenten comprender el significado semántico del esquema de anotación en lugar de simplemente aprender por asociación estadística. Dado que la producción de recursos humanos anotados con este nivel de semántica es de una alta complejidad incluso para los expertos, es poco probable que los enfoques puramente estadísticos sean suficientes para aprender en este escenario.

3.7.3. Limitaciones existentes

Durante el desarrollo de la investigación se produjo una evolución del esquema de anotación y la calidad de los corpus anotados, fundamentalmente

relacionado con el aumento de la expresividad de los conceptos compuestos. La primera versión del corpus permitía conceptos compuestos solo a través de la anotación de `Action` y sus roles correspondientes. Posteriormente se introdujo el concepto de `Predicate` y las relaciones contextuales, que permiten una representación semántica más detallada al componer conceptos complejos. Además se introdujeron las relaciones `causes` y `entails` con una semántica bien definida. Este tipo de relaciones podría permitir la construcción de sistemas de inferencia que puedan descubrir nuevos conocimientos mediante la aplicación sucesiva de reglas de inferencia.

Sin embargo, aumentar la expresividad de un esquema de anotación también introduce nuevas fuentes de ambigüedad. Durante el proceso de anotación, se evidenciaron varias fuentes de desacuerdo entre anotadores. Por ejemplo, decidir entre `Predicate` e `in-context`, o en los diferentes roles semánticos asignados a las anotaciones `target`. Uno de estos roles es similar a `MotivatedByGoal` y `UsedFor` en ConceptNet, es decir, para indicar que una `Action` se realiza con un propósito. Este uso es diferente a `causes` y `entails` y puede requerir la adición de una nueva relación semántica.

Hasta el momento, la investigación se ha centrado fundamentalmente en el idioma español, dado el predominio de los recursos en inglés en comparación con los basados en el español. Sin embargo, el esquema de anotación ha sido diseñado con el objetivo explícito de ser aplicable en muchos idiomas. Los elementos centrales son todos independientes del idioma. Esto se debe a que los conceptos, acciones, referencias y predicados, así como las relaciones semánticas definidas, se encuentran en todos los lenguajes humanos, incluso si su representación sintáctica es diferente. Con vistas a evaluar la generalización del esquema de anotación propuesto, en la edición 2021 se desarrolló una prueba de concepto para anotar un pequeño conjunto de oraciones en inglés. El éxito de esta prueba, aunque aún es incipiente, permite reconocer la potencialidad del esquema de anotación desarrollado en esta Tesis para ser generalizado a otros dominios e idiomas.

Con respecto al proceso de anotación, la principal limitación para producir recursos lingüísticos de mayor envergadura radica en el tiempo necesario para la anotación humana. En este sentido, el sistema de anotación asistida propuesto en esta investigación puede mitigar este costo, aunque también tiene limitaciones relacioandas con la complejidad de los algoritmos de aprendizaje

utilizados. El uso de modelos simples es necesario en escenarios de aprendizaje activo donde los algoritmos deben ser entrenados interactivamente, pero hay factores adicionales a considerar relacionados con la complejidad del modelo. Existe un balance interesante entre la capacidad de un modelo y su utilidad para el aprendizaje activo. Los modelos muy simples tendrán una alta incertidumbre en todas las oraciones, mientras que los modelos muy complejos sobreestimarán su propia certeza. En ambos casos, la métrica de informatividad para todas las oraciones será muy similar, lo que impide elegir las más informativas. Esto indica que puede haber un punto medio óptimo en el que el modelo aprende lo suficiente como para proporcionar sugerencias útiles y al mismo tiempo mantener un nivel adecuado de incertidumbre. Incluso en escenarios muy complejos donde los modelos del estado del arte son imposibles de entrenar interactivamente, el uso de modelos sustitutos más débiles aún puede proporcionar un beneficio significativo.

Parte II

Artículos Publicados o Aceptados

Capítulo 4

Esquema de Anotación

En este Capítulo se presenta el artículo *A General-Purpose Annotation Model for Knowledge Discovery: Case Study in Spanish Clinical Text*. Este artículo define un modelo de anotación diseñado para capturar una gran parte de la semántica del texto en lenguaje natural. Se presenta la estructura del modelo de anotación, con ejemplos de oraciones anotadas y una breve descripción de cada rol semántico y relación definida. Esta investigación se centra en una aplicación a textos clínicos en español. Sin embargo, el modelo de anotación presentado es extensible a otros dominios e idiomas. También se proporciona un ejemplo de oraciones anotadas, una guía de anotación y archivos de configuración adecuados para una herramienta de anotación usara por la comunidad científica.

Entrada bibliográfica:
Piad-Morffis, A., Guitérrez, Y., Estevez-Velarde, S., Muñoz, R. (2019, June). A general-purpose annotation model for knowledge discovery: Case study in Spanish clinical text. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 79-88).

Disponible en: http://dx.doi.org/10.18653/v1/W19-1910

Capítulo 5

Corpus eHealth-KD 2018

Ese capítulo presenta el artículo *A corpus to support eHealth Knowledge Discovery technologies*. Este artículo introduce y describe el corpus *eHealth-KD 2018*. El corpus es una colección de 1173 oraciones relacionadas con la salud en español anotadas manualmente con una estructura semántica general que captura la mayor parte del contenido, sin recurrir a etiquetas de dominio específico. La representación semántica se define e ilustra primero con oraciones de ejemplo del corpus. Se resume además el proceso de anotación y se proporcionan métricas relevantes del corpus. Finalmente, se presentan tres implementaciones computacionales, que son compatibles con los modelos de aprendizaje automático y fueron diseñadas para estimar la complejidad del aprendizaje de la semántica del corpus. El corpus resultante se utilizó como escenario de evaluación en TASS 2018 y se discuten los resultados obtenidos por los participantes. El corpus *eHealth-KD* proporciona el primer paso en el diseño de un marco semántico de propósito general que se puede utilizar para extraer conocimiento de varios dominios.

Entrada bibliográfica:
Piad-Morffis, A., Gutiérrez, Y., Muñoz, R. (2019). A corpus to support ehealth knowledge discovery technologies. *Journal of Biomedical Informatics*, 94, 103172.

Disponible en: https://doi.org/10.1016/j.jbi.2019.103172

Capítulo 6

eHealth Knowledge Discovery 2018

Ese capítulo presenta el artículo *Analysis of eHealth Knowledge Discovery Systems inthe TASS 2018 Workshop*. En este artículo se analizan los sistemas presentados en la primera edición del evento *eHealth Knowledge Discovery*. Se presenta una breve descripción de la tarea, las métricas de rendimiento, y los sistemas presentados. Además, se presenta un análisis de los resultados obtenidos por estos sistemas, enfocándose en las características de cada subtarea. Los resultados de esta primera edición del evento demostraron que el descubrimiento de conocimiento en idioma español es un área de investigación fructífera.

Entrada bibliográfica:
Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estevez-Velarde, Yudivian Almeida-Cruz, Andrés Montoyo, and Rafael Muñoz. Analysis of eHealth Knowledge Discovery Systems in the TASS 2018 Workshop. *Procesamiento del Lenguaje Natural*, 62(0):13–20, 2019. ISSN 1989-7553.

Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5947.

Capítulo 7

Ecosistema Computacional

Este capítulo presenta el artículo *A Computational Ecosystem to Support eHealth Knowledge Discovery Technologies in Spanish*. Este trabajo describe un ecosistema que facilita la investigación y el desarrollo en el descubrimiento del conocimiento en el dominio biomédico, específicamente en el idioma español. Con este fin, se desarrollan y comparten varios recursos con la comunidad de investigación, incluido un nuevo modelo de anotación semántica, un corpus anotado de 1.045 oraciones y recursos computacionales para construir y evaluar técnicas automáticas de descubrimiento de conocimiento. Además, se define una tarea de aprendizaje automático con criterios de evaluación objetivos, y se diseña un entorno de evaluación en línea, lo que permite a los investigadores interesados en esta tarea obtener resultados inmediatos y compararlos con el estado del arte. Como caso de estudio, se analizan los resultados de un evento competitivo basado en estos recursos y se brindan pautas para futuras investigaciones.

Entrada bibliográfica:
Piad-Morffis, A., Gutiérrez, Y., Almeida-Cruz, Y., Muñoz, R. (2020). A computational ecosystem to support eHealth Knowledge Discovery technologies in Spanish. *Journal of Biomedical Informatics*, 109, 103517.

Disponible en: https://doi.org/10.1016/j.jbi.2020.103517

Capítulo 8

eHealth Knowledge Discovery 2021

Este capítulo presenta el artículo *Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021*. Este artículo resume los resultados de la última edición del evento *eHealth Knowledge Discovery*. A partir de una comparación con los resultados de ediciones anteriores, permite apreciar el avance que se ha producido en el campo en los cuatro años en los que se ha organizado el evento. En esta edición la tarea se extendió a nuevos dominios e idiomas, lo que permitió contrastar los resultados de sistemas de dominio específico con respecto a sistemas diseñados para la generalización. Estos resultados demuestran que la tarea de descubrimiento de conocimiento en lenguaje natural en múltiples dominios aún representa un reto científico y tecnológico relevante.

Entrada bibliográfica:
Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estevez-Velarde, Yudivian Almeida-Cruz, Andrés Montoyo, and Rafael Muñoz. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021.*Procesamiento del Lenguaje Natural*, 67(0):233–242. 2021. ISSN 1989-7553.

Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392

Capítulo 9

Anotación Asistida

Este capítulo presenta el artículo *Active Learning for Assisted Corpus Construction:A Case Study in Knowledge Discovery from Biomedical Text.* Este artículo define un enfoque de aprendizaje activo que tiene como objetivo reducir el esfuerzo humano requerido durante la anotación de corpus en lenguaje natural, compuestos por entidades y relaciones semánticas. Nuestro enfoque ayuda a los anotadores humanos seleccionando inteligentemente las oraciones más informativas para anotar y pre-anotándolas con algunas entidades y relaciones semánticas. Se define una estrategia basada en la incertidumbre con un factor de densidad ponderado, utilizando métricas de similitud basadas en *embeddings* de oraciones. Los resultados experimentales sugieren que la estrategia de consulta reduce entre un 35 % y 40 % el número de oraciones que se deben anotar manualmente para desarrollar sistemas capaces de alcanzar un resultado objetivo de F_1, mientras que la estrategia de pre-anotación produce una reducción adicional de 24 % en el tiempo total de anotación.

Entrada bibliográfica:
Hian Cañizares-Díaz, Alejandro Piad-Morffis, Suilan Estevez-Velarde, Yoan Gutiérrez, Yudivián Almeida Cruz, Andrés Montoyo and Rafael Muñoz-Guillena. Active Learning for Assisted Corpus Construction: A Case Study in Knowledge Discovery from Biomedical Text. *Proceedings of RANLP 2021.*

Disponible en: https://ranlp.org/ranlp2021/proceedings-20Sep.pdf

Parte III

Conclusiones y Recomendaciones

Capítulo 10

Conclusiones

El volumen de información que se publica cada día es una fuente de conocimiento relevante de gran utilidad para la comunidad científica. A partir del análisis de múltiples fuentes es posible encontrar conexiones que permitan potencialmente descubrir nuevo conocimiento que se encuentra disperso en las redes. A pesar de esto, el alto número de fuentes a revisar es imposible de procesar por humanos, lo que dificulta este proceso de descubrimiento de conocimiento. En este contexto cobra especial relevancia el desarrollo de técnicas automáticas para la extracción y descubrimiento de este conocimiento.

El dominio de la salud es uno de los que más puede beneficiarse del descubrimiento automático de conocimiento, debido a la velocidad y el volumen con el que se publican nuevos resultados. La reciente pandemia de COVID-19 es solo un ejemplo más de la necesidad de conectar de forma automática los hechos y resultados publicados a diario. Para construir sistemas computacionales capaces de realizar esta tarea, es necesario no solo la creación de recursos lingüísticos que sirvan para los modelos de aprendizaje automático, sino también la creación de una infraestructura que permita a los investigadores evaluar sus sistemas de forma efectiva y compararlos con el estado del arte.

Esta investigación presenta el diseño y la construcción de un ecosistema para el desarrollo de tecnologías de descubrimiento de conocimiento, enfocado en el dominio de la salud y el idioma español. Este ecosistema incluye

recursos lingüísticos, herramientas computacionales y una metodología para la evaluación de nuevos enfoques.

Con el objetivo de representar de forma computacionalmente manejable el lenguaje natural, se definió un modelo de anotación para capturar el contenido semántico más relevante de una oración, basado en una estructura Sujeto-Acción-Objetivo y relaciones semánticas adicionales. El modelo no incluye entidades o relaciones de dominio específico para ser lo más general posible. Este modelo está diseñado para ser suficientemente expresivo pero que aún así sea factible tanto su anotación por expertos humanos con un alto grado de acuerdo entre diferentes anotadores como la aplicación de técnicas de aprendizaje automático.

Basado en este esquema de anotación, se anotaron manualmente a lo largo de la investigación cuatro versiones de un corpus con un total de 41,163 elementos semánticos repartidos en 3,068 oraciones, tomando como base información en el dominio de salud en idioma español. A pesar del enfoque en el dominio de la salud, a modo de caso de estudio se incluyen 300 oraciones en un dominio alternativo (reportajes periodísticos), 200 oraciones en idioma inglés, que demuestran la capacidad de generalización del esquema de anotación. Los cuatro versiones de un corpus fueron anotados por equipos compuestos por anotadores expertos y no expertos, siguiendo un protocolo que garantizó un alto grado de acuerdo en los elementos anotados. El propósito de estos recursos lingüísticos es permitir la creación de sistemas de descubrimiento de conocimiento a partir del uso de técnicas de aprendizaje automático. Con este propósito, se ha organizado una campaña de evaluación, *eHealth-KD challenge*, de la que se han producido cuatro ediciones consecutivas. En total han participado un total de 33 equipos de investigadores de diferentes nacionalidades con múltiples estrategias, principalmente enfocadas en arquitecturas de aprendizaje profundo. Para permitir una comparación entre diferentes sistemas, se diseñó un escenario que consiste de dos subtareas fundamentales, y se definieron un conjunto de métricas para su evaluación. Los resultados obtenidos en las cuatro ediciones de esta campaña demuestran que el reto de descubrir conocimiento automáticamente en lenguaje natural es un problema aún abierto pero se ha producido una mejora considerable en la efectividad de los sistemas presentados de una edición a la siguiente, fomentado por el reciente desarrollo de nuevas arquitecturas de aprendizaje profundo para el procesamiento de lenguaje natural.

Más allá de las ediciones de la campaña *eHealth-KD*, esta investigación se propuso construir un ecosistema de evaluación continua que pueda ser utilizado de forma independiente por cualquier investigador para evaluar nuevos enfoques y compararse con el estado del arte. Este ecosistema se pone a disposición de la comunidad científica, que consiste una infraestructura y un conjunto de herramientas (incluyendo implementaciones básicas con las que compararse), un entorno de evaluación continua en la nube, y estadísticas actualizadas sobre el estado del arte de la tarea *eHealth-KD*. Todos estos recursos, incluyendo los corpus anotados y las herramientas computacionales están publicados bajo licencias de código abiertas y disponibles para su uso en línea[1].

Finalmente, a partir de las experiencias acumuladas durante el proceso de anotación, se implementó un sistema de anotación asistida que permite reducir considerablemente el costo manual de producir nuevos recursos lingüísticos en esta área. Este sistema utiliza técnicas de aprendizaje automático para seleccionar las oraciones más útiles a anotar, y garantizar un balance de los datos en un corpus con el menor esfuerzo posible por parte de los anotadores humanos. Esta herramienta está basada en componentes de código abierto y se encuentra disponible para el uso de la comunidad científica[2].

Todos los resultados presentados en esta Tesis son parte de una línea de investigación activa para aprovechar la semántica de propósito general y las tecnologías basadas en el conocimiento junto con las nuevas arquitecturas de aprendizaje profundo en la construcción de tecnologías automáticas de descubrimiento de conocimiento. Como parte de esta línea, se continúa trabajando en la expansión de los recursos lingüísticos creados y en la organización de campañas de evaluación para fomentar la investigación en esta temática en la comunidad, así como en el desarrollo de nuevas herramientas computacionales basadas en técnicas de aprendizaje automático para el descubrimiento de conocimiento a partir de lenguaje natural.

[1] https://ehealthkd.github.io

[2] https://github.com/knowledge-learning/assisted-annotation

10.1. Publicaciones

Los resultados de investigación parciales que conforman esta tesis se han publicado anteriormente en diversos artículos. A continuación se listan aquellas publicaciones relacionadas con esta investigación en las que ha participado el autor:

- Overview of TASS 2018: Opinions, health and emotions [168]
- Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019 [169]
- Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020 [170]
- Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021 [171]
- A general-purpose annotation model for knowledge discovery: Case study in Spanish clinical text [172]
- A corpus to support eHealth Knowledge Discovery technologies [173]
- A Computational Ecosystem to Support eHealth Knowledge Discovery Technologies in Spanish [174]
- Gathering object interactions as semantic knowledge [175]
- TASS 2018: The strength of deep learning in language understanding tasks [176]
- A Neural Network Component for Knowledge-Based Semantic Representations of Text [177]
- Analysis of eHealth knowledge discovery systems in the TASS 2018 workshop [178]
- Active Learning for Assisted Corpus Construction:A Case Study in Knowledge Discovery from Biomedical Text [179]
- Knowledge Discovery in COVID-19 Research Literature [180]

Adicionalmente, se listan a continuación las ediciones del evento *eHealth Knowledge Discovery* organizadas como parte del marco de evaluación de esta investigación:

- eHealth KD 2018 como Workshop en el evento TASS 2018
  `http://www.sepln.org/workshops/tass/2018/task-3`
- eHealth KD 2019 como Workshop en el evento IBERLEF 2019
  `https://knowledge-learning.github.io/ehealthkd-2019`
- eHealth KD 2020 como Workshop en el evento IBERLEF 2020
  `https://knowledge-learning.github.io/ehealthkd-2020`
- eHealth KD 2021 como Workshop en el evento IBERLEF 2021
  `https://ehealthkd.github.io/2021`

Capítulo 11

Trabajo Futuro

La investigación desarrollada en esta tesis puede ser continuada desde múltiples perspectivas. En primer lugar, los modelos teóricos de representación del conocimiento diseñados así como los algoritmos presentados, pueden ser mejorados y adaptados a nuevos escenarios. Por otro lado, el ecosistema diseñado permite continuar con la organización de eventos competitivos en diferentes tareas y dominios para incentivar el desarrollo de nuevas técnicas de descubrimiento de conocimiento. En este capítulo se presentan las ideas que consideramos más relevantes y que pueden proveer un mayor aporte a los resultados y aplicaciones futuras de esta investigación.

Desde la arista de desarrollo y creación de corpus, una de las líneas de investigación más claras es la anotación de nuevos recursos lingüísticos en otros dominios e idiomas. Entre los dominios más interesantes a analizar se encuentran los artículos científicos, las noticias y los artículos enciclopédicos. Estas son fuentes de información factual que presentan diferentes características sintácticas pero que potencialmente pueden ser capturadas con las mismas estructuras semánticas definidas en esta investigación. Anotar nuevos dominios permitiría también ampliar el *eHealth-KD* a una audiencia mayor, por ejemplo, los investigadores interesados en el fenómeno de las noticias falsas. Además, esto abre las puertas a la definición de tareas multi-dominio, donde los mismos sistemas sean evaluados en corpus de dominios diferentes a los que fueron entrenados, requiriendo el desarrollo de técnicas de transferencia de aprendizaje automático.

En la dirección de modernizar la tarea *eHealth-KD* hay dos escenarios nuevos interesantes de explorar. En primer lugar, adicionar el problema de reconocimiento de los atributos, que actualmente son anotados en el corpus pero no evaluados en la tarea. Este escenario ya introduce algunos problemas computacionales interesantes, por ejemplo, la identificación de la negación y su ámbito. Con vistas a avanzar hacia una representación semántica de más alto nivel, el otro escenario de interés consiste en la normalización de las entidades extraídas en función de bases de datos estructuradas, por ejemplo, Wikidata. Estas adiciones complejizan la tarea *eHealth-KD*, pero como mismo se separan la extracción de entidades y relaciones en diferentes subtareas, se pueden definir subtareas específicas para estos problemas permitiendo que investigadores especializados en un área participen en aquellos problemas de su interés.

El esquema de anotación propuesto en esta investigación ha demostrado ser capaz de capturar una porción significativa de la semántica del lenguaje natural. Sin embargo, aún mantiene algunas inconsistencias y ambigüedades que pueden ser mejoradas con pocas modificaciones. Una de las principales fuentes de ambigüedad identificada es la similitud entre los roles de `Predicate` y la relación `in-context`. Es posible que una de estas variantes sea redundante, o que deba definirse de manera más clara la diferencia entre ambos patrones de anotación. La adición más importante al esquema de anotación consistiría en un nuevo tipo de relación semántica que permita denotar el propósito de una acción o evento.

Finalmente, la estrategia de aprendizaje activo desarrollada en la investigación brinda una solución prometedora al problema del costo de la anotación manual, pero solo ha sido evaluada en escenarios simulados. La propuesta más relevante en este sentido consiste en aplicar esta estrategia para la anotación de los recursos lingüísticos de la próxima edición del *eHealth-KD*. Para esto es necesario adaptar la propuesta a un escenario con múltiples anotadores por cada oración. En este escenario, el propio desacuerdo entre los anotadores introduce una nueva fuente de incertidumbre que puede ser utilizada para optimizar el proceso de anotación.

Las ideas sugeridas en este capítulo son solo algunas de las posibilidades más claras para continuar el desarrollo de las técnicas y recursos propuestos en esta tesis. El trabajo desarrollado forma parte de una línea de investigación

activa. Estas y otras ideas serán tenidas en cuenta para estimular aún más el desarrollo de tecnologías de descubrimiento de conocimiento, siempre desde la perspectiva de aprovechar el esfuerzo humano y los recursos computacionales en estrecha colaboración. Visualizamos un futuro donde los seres humanos y las computadoras trabajen juntos en la solución de los problemas más apremiantes, cada uno poniendo sus mejores habilidades al servicio de las futuras generaciones.

Bibliografía

[1] 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.

[2] Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook. 2005.

[3] 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.

[4] 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.

[5] 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.

[6] 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 Multilingual-BIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.)*, 2018.

[7] 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.

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

[9] 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.

[10] 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.

[11] Isabel Moreno, Ester Boldrini, Paloma Moreda, and M. Teresa Román-Ferri. Drugsemantics: A corpus for named entity recognition in spanish summaries of product characteristics. *Journal of Biomedical Informatics*, 72:8–22, 2017. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.06.013. URL https://www.sciencedirect.com/science/article/pii/S1532046417301363.

[12] 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.

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

[14] 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.

[15] 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.

[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] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Saint John Walker. Big data: A revolution that will transform how we live, work, and think, 2014.

[25] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: from big data to big impact. *MIS quarterly*, pages 1165–1188, 2012.

[26] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

[27] Dhavan V Shah, Joseph N Cappella, and W Russell Neuman. Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1):6–13, 2015.

[28] Nicola Guarino. Formal ontology, conceptual analysis and knowledge representation. *International journal of human-computer studies*, 43 (5-6):625–640, 1995.

[29] Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, 2006.

[30] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos. Ontology population and enrichment:

State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution*, pages 134–166. Springer-Verlag, 2011.

[31] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12, 2005.

[32] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. *Commun. ACM*, 61(5):103–115, April 2018. ISSN 0001-0782. doi: 10.1145/3191513. URL http://doi.acm.org/10.1145/3191513.

[33] Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. *EmotiNet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories*, pages 27–39. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-22327-3. doi: 10.1007/978-3-642-22327-3_4. URL http://dx.doi.org/10.1007/978-3-642-22327-3_4.

[34] Tom M Mitchell et al. Machine learning. 1997.

[35] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

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

[37] John F Sowa et al. *Knowledge representation: logical, philosophical, and computational foundations*, volume 13. Brooks/Cole Pacific Grove, 2000.

[38] Oscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez. Methodologies, tools and languages for building ontologies. where is their meeting point? *Data & Knowledge Engineering*, 46(1):41 – 64, 2003. ISSN 0169-023X. doi: https://doi.org/10.1016/S0169-023X(02)00195-7. URL http://www.sciencedirect.com/science/article/pii/S0169023X02001957.

[39] Ronald J Brachman, Hector J Levesque, and Raymond Reiter. *Knowledge representation*. MIT press, 1992.

[40] Michel Chein and Marie-Laure Mugnier. *Graph-based knowledge representation: computational foundations of conceptual graphs*. Springer Science & Business Media, 2008.

[41] Wilson Wong, Wei Liu, and Mohammed Bennamoun. Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44(4):20, 2012.

[42] Dehai Zhang, Naiyao Wang, Ye Yuan, Bin Wang, and Yun Yang. Fuzzy ontology induction in the cognitive model of ontology learning. In *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on*, pages 739–744. IEEE, 2017.

[43] Fernando Bobillo and Umberto Straccia. Fuzzy ontology representation using owl 2. *International Journal of Approximate Reasoning*, 52(7): 1073–1094, 2011.

[44] Alan L Rector, JE Rogers, Pieter E Zanstra, and Egbert Van Der Haring. Opengalen: open source medical terminology and tools. In *AMIA Annual Symposium Proceedings*, volume 2003, page 982. American Medical Informatics Association, 2003.

[45] Gene Ontology Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

[46] P. Wongthongtham, E. Chang, T. Dillon, and I. Sommerville. Development of a software engineering ontology for multisite software development. *IEEE Transactions on Knowledge and Data Engineering*, 21(8): 1205–1217, Aug 2009. ISSN 1041-4347. doi: 10.1109/TKDE.2008.209.

[47] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[48] Pablo N Mendes, Max Jakob, and Christian Bizer. Dbpedia: A multilingual cross-domain knowledge base. In *LREC*, pages 1813–1817, 2012.

[49] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227, 2009.

[50] Steffen Staab and Rudi Studer. *Handbook on ontologies*. Springer Science & Business Media, 2010.

[51] Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37-38:132 – 151, 2016. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2015.12.004. URL http://www.sciencedirect.com/science/article/pii/S1570826815001456.

[52] Philipp Cimiano, Alexander Mädche, Steffen Staab, and Johanna Völker. Ontology learning. In *Handbook on ontologies*, pages 245–267. Springer, 2009.

[53] Ken Barker, Bhalchandra Agashe, Shaw Yi Chaw, James Fan, Noah Friedland, Michael Glass, Jerry Hobbs, Eduard Hovy, David Israel, Doo Soon Kim, et al. Learning by reading: A prototype system, performance baseline and lessons learned. In *AAAI*, volume 7, pages 280–286, 2007.

[54] Philipp Cimiano and Johanna Völker. text2onto. In *International Conference on Application of Natural Language to Information Systems*, pages 227–238. Springer, 2005.

[55] Paul Buitelaar and Michael Sintek. Ontolt version 1.0: Middleware for ontology extraction from text. In *Proc. of the Demo Session at the International Semantic Web Conference*, 2004.

[56] Euthymios Drymonas, Kalliopi Zervanou, and Euripides GM Petrakis. Unsupervised ontology acquisition from plain texts: the OntoGain system. In *International Conference on Application of Natural Language to Information Systems*, pages 277–287. Springer, 2010.

[57] Udo Hahn and Martin Romacker. The syndikate text knowledge base generator. In *Proceedings of the first international conference on Human*

language technology research, pages 1–6. Association for Computational Linguistics, 2001.

[58] Harith Alani, Sanghee Kim, David E Millard, Mark J Weal, Wendy Hall, Paul H Lewis, and Nigel R Shadbolt. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.

[59] Paul Buitelaar, Philipp Cimiano, Stefania Racioppa, and Melanie Siegel. Ontology-based information extraction with soba. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2006.

[60] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. Learning to construct knowledge bases from the world wide web. *Artificial intelligence*, 118 (1-2):69–113, 2000.

[61] Heiko Stoermer, Ignazio Palmisano, Domenico Redavid, Luigi Iannone, Paolo Bouquet, and Giovanni Semeraro. *Contextualization of a RDF Knowledge Base in the VIKEF Project*, pages 101–110. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-49377-8. doi: 10.1007/11931584_13. URL https://doi.org/10.1007/11931584_13.

[62] Christopher Brewster, Fabio Ciravegna, and Yorick Wilks. User-centred ontology learning for knowledge management. *Natural Language Processing and Information Systems*, pages 203–207, 2002.

[63] Sang-Soo Kim, Jeong-Woo Son, Seong-Bae Park, Se-Young Park, Changki Lee, Ji-Hyun Wang, Myung-Gil Jang, and Hyung-Geun Park. Optima: An ontology population system. In *3rd Workshop on Ontology Learning and Population (July 2008)*, 2008.

[64] Nicolas Weber and Paul Buitelaar. Web-based ontology learning with isolde. In *Proc. of the Workshop on Web Content Mining with Human Language at the International Semantic Web Conference, Athens GA, USA*, volume 11, 2006.

[65] Fabian M Suchanek, Georgiana Ifrim, and Gerhard Weikum. Leila: Learning to extract information by linguistic analysis. In *Proceedings*

of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pages 18–25, 2006.

[66] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110. ACM, 2004.

[67] David Faure and Thierry Poibeau. First experiments of using semantic knowledge learned by asium for information extraction task using intex. In *Proceedings of the ECAI workshop on Ontology Learning*, 2000.

[68] Silvana Castano, Sofia Espinosa, Alfio Ferrara, Vangelis Karkaletsis, Atila Kaya, Sylvia Melzer, Ralf Möller, Stefano Montanelli, and Georgios Petasis. Ontology dynamics with multimedia information: The boemie evolution methodology. In *International Workshop on Ontology Dynamics (IWOD-07)*, page 41, 2007.

[69] Oded Maimon and Abel Browarnik. Ontology Learning from Text: Why the Ontology Learning Layer Cake is Not Viable. *Int. J. Signs Semiot. Syst.*, 4(2):1–14, July 2015. ISSN 2155-5028. doi: 10.4018/IJSSS.2015070101. URL http://dx.doi.org/10.4018/IJSSS.2015070101.

[70] Chris Alberti, Daniel Andor, Ivan Bogatyy, Michael Collins, Dan Gillick, Lingpeng Kong, Terry Koo, Ji Ma, Mark Omernick, Slav Petrov, Chayut Thanapirom, Zora Tung, and David Weiss. Syntaxnet models for the conll 2017 shared task. *CoRR*, abs/1703.04929, 2017. URL http://arxiv.org/abs/1703.04929.

[71] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

[72] Janez Brank, Dunja Mladenic, and Marko Grobelnik. Gold standard

based ontology evaluation using instance assignment. In *Workshop on Evaluation of Ontologies for the Web, EON*. Edinburgh, UK, 2006.

[73] Adolfo Lozano-Tello and Asunción Gómez-Pérez. Ontometric: A method to choose the appropriate ontology. *Journal of database management*, 2(15):1–18, 2004.

[74] Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. Frame-based ontology population with pikes. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3261–3275, 2016.

[75] Iryna Gurevych, Rainer Malaka, Robert Porzel, and Hans-Peter Zorn. Semantic coherence scoring using an ontology. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 9–16. Association for Computational Linguistics, 2003.

[76] Janez Brank, Marko Grobelnik, and Dunja Mladenić. A survey of ontology evaluation techniques. 2005.

[77] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data driven ontology evaluation. 2004.

[78] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

[79] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3): 37, 1996. doi: 10.1609/aimag.v17i3.1230. URL https://doi.org/10.1609/aimag.v17i3.1230.

[80] 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. doi: 10.1002/widm.1093. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1093.

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

doi: 10.1007/978-3-319-46448-0_51. URL https://doi.org/10.1007%2F978-3-319-46448-0_51.

[82] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 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, 2010. URL http://dl.acm.org/citation.cfm?id=2898607.2898816.

[83] Matthew S. Simpson and Dina Demner-Fushman. *Biomedical Text Mining: A Survey of Recent Progress*, pages 465–517. Springer US, Boston, MA, 2012. ISBN 978-1-4614-3223-4. doi: 10.1007/978-1-4614-3223-4_14. URL https://doi.org/10.1007/978-1-4614-3223-4_14.

[84] 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, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-2322.

[85] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998. doi: 10.3115/980451.980860. URL http://dx.doi.org/10.3115/980451.980860.

[86] S. Estevez-Velarde, Y. Gutierrez, A. Montoyo, A. Piad-Morffis, R. Munoz, and Y. Almeida-Cruz. Gathering object interactions as semantic knowledge (accepted). In *Proceedings of the 2017 International Conference on Artificial Intelligence (ICAI'17)*, 2018.

[87] Thomas C Rindflesch and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of biomedical informatics*, 36(6):462–477, 2003.

[88] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1): D267–D270, 2004. doi: 10.1093/nar/gkh061. URL http://dx.doi.org/10.1093/nar/gkh061.

[89] George Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[90] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2): 167–195, 2015.

[91] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686, 2012.

[92] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. Overview of the clef ehealth evaluation lab 2016. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 255–266. Springer, 2016.

[93] Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. Background to framenet. *International journal of lexicography*, 16(3): 235–250, 2003.

[94] Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, Philadelphia, PA, USA, 2005. URL https://repository.upenn.edu/dissertations/AAI3179808. AAI3179808.

[95] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[96] James Pustejovsky. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10, 2013.

[97] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending verbnet with novel verb classes. In *LREC*, pages 1027–1032, 2006.

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

[99] Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

[100] Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. Biomedical event extraction using abstract meaning representation. *BioNLP 2017*, pages 126–135, 2017. doi: 10.18653/v1/W17-2315. URL http://dx.doi.org/10.18653/v1/W17-2315.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

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

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

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

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

[111] 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.

[112] 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.

[113] Diana Maynard Hamish Cunningham and Kalina Bontcheva. *Text Processing with GATE (Version 6).* University of Sheffield D, 2011.

[114] Mark A Musen. The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12, 2015.

[115] Eva Mújdricza-Maydt, Silvana Hartmann, Iryna Gurevych, and Anette Frank. Combining semantic annotation of word sense & semantic roles: A novel annotation scheme for verbnet roles on german language data.

In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3031–3038, 2016.

[116] Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 8. using framenet for the semantic analysis of german: Annotation, representation, and automation. In *Multilingual FrameNets in computational lexicography*, pages 209–244. De Gruyter Mouton, 2009.

[117] Tjong Kim Sang. Ef and de meulder, f.(2003). In *Introduction to the CoNLL-2003 shared task: Language independent named entity recognition. InProceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 142–147, 2003.

[118] Jin-Dong Kim and Yue Wang. Pubannotation-a persistent and sharable corpus and annotation repository. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205, 2012.

[119] György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, 2008.

[120] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[121] Burr Settles. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14): 3191–3192, 2005.

[122] Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17, 2010.

[123] Cecilia N Arighi, Zhiyong Lu, Martin Krallinger, Kevin B Cohen, W John Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy H Wu. Overview of the biocreative iii workshop. *BMC bioinformatics*, 12 (8):1–9, 2011.

[124] 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.

[125] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences."n the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer, 2002.

[126] Erik Cambria, Robyn Speer, Catherine Havasi, and Amir Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*, 2010.

[127] Sven J Körner and Torben Brumm. Natural language specification improvement with ontologies. *International Journal of Semantic Computing*, 3(04):445–470, 2009.

[128] Robyn Speer, Catherine Havasi, K Nichole Treadway, and Henry Lieberman. Finding your way in a multi-dimensional semantic space with luminoso. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 385–388, 2010.

[129] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30 (1-7):107–117, 1998.

[130] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[131] David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. Watson: beyond jeopardy! *Artificial Intelligence*, 199: 93–105, 2013.

[132] Balakrishnan Chandrasekaran. Generic tasks in knowledge-based reasoning: High-level building blocks for expert system design. *IEEE expert*, 1(3):23–30, 1986.

[133] Murphy Kevin. Machine learning: a probabilistic perspective, 2012.

[134] Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[135] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[136] Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, and Eric Xing. Entity hierarchy embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1292–1300, 2015.

[137] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[138] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[139] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[140] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

[141] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[142] Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1-2):135–150, 2002.

[143] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

[144] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[145] David Cohn. Active Learning. In *Encyclopedia of Machine Learning*, 2010.

[146] H. S. Seung, M. Opper, and H. Sompolinsky. Query by Committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130417. URL https://doi.org/10.1145/130385.130417.

[147] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In *ICML*, 2017.

[148] Janez Kranjc, Jasmina Smailovic, Vid Podpecan, Miha Grcar, Martin Znidarsic, and Nada Lavrac. Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Inf. Process. Manag.*, 51:187–203, 2015.

[149] Hong-Tao Zhang, Min-Lie Huang, and Xiao-Yan Zhu. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology*, 27(6):1302–1313, 2012.

[150] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148 – 156. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. doi: https://doi.org/10.1016/B978-1-55860-335-6.50026-X. URL http://www.sciencedirect.com/science/article/pii/B978155860335650026X.

[151] Burr Settles, Mark Craven, and Soumya Ray. Multiple-Instance Active Learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis,

editors, *Advances in Neural Information Processing Systems 20*, pages 1289–1296. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3252-multiple-instance-active-learning.pdf.

[152] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[153] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

[154] Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. 2009.

[155] Roselyne Tchoua, Aswathy Ajith, Zhi Hong, Logan T. Ward, Kyle Chard, Debra Audus, Shrayesh Patel, Juan de Pablo, and Ian T Foster. Active Learning Yields Better Training Data for Scientific Named Entity Recognition. *2019 15th International Conference on eScience (eScience)*, pages 126–135, 2019.

[156] Venkata Vamsikrishna Meduri, Lucian Popa, Prithviraj Sen, and Mohamed Sarwat. A Comprehensive Benchmark Framework for Active Learning Methods in Entity Matching. *ArXiv*, abs/2003.13114, 2020.

[157] Vincent Claveau and Ewa Kijak. Strategies to Select Examples for Active Learning with Conditional Random Fields. In *CICLing*, 2017.

[158] Bill Yuchen Lin, Dongho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren. AlpacaTag: An Active Learning-based Crowd Annotation Framework for Sequence Tagging. In *ACL*, 2019.

[159] Lisheng Fu and Ralph Grishman. An efficient active learning framework for new relation types. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 692–698, 2013.

[160] Ning Gao, Nikos Karampatziakis, Rahul Potharaju, and Silviu Cucerzan. Active Entity Recognition in Low Resource Settings. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019.

[161] Rafael Glauber. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. *CEUR Proceedings*, 2019.

[162] 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.

[163] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[164] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[165] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

[166] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[167] 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.

[168] 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 Munoz, et al. Overview of tass 2018: Opinions, health and emotions. 2018.

[169] Alejandro Piad-Morffis, Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Munoz, 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.

[170] Alejandro Piad-Morffis, Yoan Gutiérrez, Hian Cañizares-Diaz, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, and Andrés Montoyo. Overview of the ehealth knowledge discovery challenge at iberlef 2019. In *IberLEF@ SEPLN*, pages 1–16, 2020.

[171] Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estévez-Velarde, Yudivián Almeida-Cruz, Andrés Montoyo, and Rafael Muñoz. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021. *Procesamiento del Lenguaje Natural*, To appear, 2021.

[172] Alejandro Piad-Morffis, Yoan Guitérrez, Suilan Estevez-Velarde, and Rafael Muñoz. A general-purpose annotation model for knowledge discovery: Case study in spanish clinical text. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 79–88, 2019.

[173] Alejandro Piad-Morffis, Yoan Gutiérrez, and Rafael Muñoz. A corpus to support ehealth knowledge discovery technologies. *Journal of biomedical informatics*, 94:103172, 2019.

[174] Alejandro Piad-Morffis, Yoan Gutiérrez, Yudivian Almeida-Cruz, and Rafael Muñoz. A computational ecosystem to support ehealth knowledge discovery technologies in spanish. *Journal of biomedical informatics*, 109:103517, 2020.

[175] Suilan Estevez-Velarde, Yoan Gutierrez, Andres Montoyo, Alejandro Piad-Morffis, Rafael Munoz, and Yudivian Almeida-Cruz. Gathering object interactions as semantic knowledge. In *International Conference on Artificial Intelligence*, 2018.

[176] 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. *CEUR Proceedings*, 2018.

[177] Alejandro Piad-Morffis, Rafael Muñoz, Yoan Gutiérrez, Yudivian Almeida-Cruz, Suilan Estevez-Velarde, and Andrés Montoyo. A neural network component for knowledge-based semantic representations of text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 904–911, 2019.

[178] Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estévez-Velarde, Yudivián Almeida-Cruz, Andrés Montoyo, and Rafael Muñoz. Analysis of ehealth knowledge discovery systems in the tass 2018 workshop. *Procesamiento del Lenguaje Natural*, 62(0):13–20, 2019.

[179] Hian Cañizares Díaz, Alejandro Piad-Morffis, Rafael Muñoz, Yoan Gutiérrez, Yudivian Almeida-Cruz, Suilan Estevez-Velarde, and Andrés Montoyo. Active Learning for Assisted Corpus Construction:A Case Study in Knowledge Discovery from Biomedical Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021.

[180] Ernesto Estevanell-Valladares, Alejandro Piad-Morffis, Rafael Muñoz, Yoan Gutiérrez, Yudivian Almeida-Cruz, Suilan Estevez-Velarde, and Andrés Montoyo. Knowledge Discovery in COVID-19 Research Literature. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021.