

Tesis

Maestría en Bioinformática

PEDECIBA

Identificación de genes vinculados al diagnóstico a partir de la información bibliográfica disponible y la historia clínica

Ing. Fernando López Bello (Facultad de Ingeniería - UdelaR)

Tutores

Dr. Hugo Naya (Unidad de Bioinformática - Institut Pasteur de Montevideo)

Dr. Víctor Raggio (Departamento de Genética - Facultad de Medicina - UdelaR)

Dra. Aiala Rosá (Instituto de Computación - Facultad de Ingeniería - UdelaR)

Tribunal

Dr. José Badano (Institut Pasteur de Montevideo)

Dr. Horacio Botti (Facultad de Medicina, UDeLaR)

Dra. Dina Wonsever (Facultad de Ingeniería, UDeLaR)

Mayo 2019

Montevideo, Uruguay

Contenido

Contenido.....	2
Agradecimientos	3
Resumen	4
1. Introducción.....	5
2. Recursos y Métodos.....	9
2.1 Recursos.....	9
2.2 Métodos.....	14
2.2.1 Procesamiento de Abstracts	14
2.2.2 Procesamiento de Resúmenes de Historias Clínicas.....	15
2.3 Notas de Implementación.....	16
3. Artículo.....	19
4. Aplicación de Supervisión Distante.....	31
4.1 Contexto.....	31
4.2 Estrategia	31
4.3 Implementación	32
4.4 Preparación de Datos para Entrenamiento	34
Métodos de Aprendizaje Automático	35
4.5 Resultados.....	37
4.5.1 Entrenamiento para Diferentes Métodos de Aprendizaje Automático.....	37
4.5.2 Incidencia en la Performance de Extracción de Relaciones.....	38
5. Consideraciones Finales.....	41
Bibliografía	43

Agradecimientos

The biggest emotion in creation is the bridge to optimism.

— *Brian May, guitarrista de Queen a los 22,
doctorado en Astrofísica a los 60.*

*Este trabajo está dedicado a quienes más me tuvieron paciencia
y prestaron su tiempo: Camilo, Felipe, Gabriela.*

No hay dedicatoria suficiente para devolver el aliento que tuve desde adentro de casa, sobre todo a mi compañera, Gabi.

Gracias por la enorme paciencia de los tutores en este largo camino: Aiala, Hugo, Víctor. A la Dra. Laura Rodríguez por su aporte en las evaluaciones de resultados.

A Oscar, primer instigador de esta aventura, y a la familia en Bahía Blanca y Barcelona, a quienes nunca logré explicar bien de qué iba todo esto. A mis divinos amigos, Serrat dixit, “unos atorrantes, sueños imprevistos que buscan sus piedras filosofales”.

A profesores inspiradores a lo largo de estos años de estudio: Gustavo Guerberoff (Probabilidad y Estadística), Héctor Musto (Genética y Evolución), Ricardo Fraiman (Data Mining), Margot Paulino (Bioinformática Estructural), Guillermo Moncecchi (Lenguaje Natural).

A los compañeros de Quanam que alentaron, criticaron, elogiaron e hicieron maquetas a partir de este trabajo: Gonzalo Herrera, Luis Vázquez, Guillermo Spinelli, Marcelo Acerenza, Leonardo Loureiro, Edgardo Noya, Fabricio Gabrielli, Mayari Arruabarrena, Milena Brum, Miguel Rojas, Brandon Assandri.

Thanks to Killian Thiel from KNIME GmbH - Berlin, Germany, for facilitating academic software licensing.

El uso de SNOMED CT fue posible gracias al licenciamiento cedido por Salud.uy, sobre SNOMED CT edición Uruguay.

Gracias especialmente al apoyo de ICT4V (ict4v.org) para hacer posible este trabajo y la publicación del artículo.

Those who can imagine anything, can create the impossible.

— *Alan Turing, padre de la computación e inteligencia artificial.*

Resumen

Este trabajo propone aprovechar las tecnologías de procesamiento del lenguaje natural desde el punto de vista de los campos de investigación y diagnóstico médico y genético, y más en general, la biotecnología y la bioinformática, como contribución al desarrollo de la aplicación médica de los análisis genómicos basados en Secuenciación de Próxima Generación. En particular, buscamos desarrollar un marco que permita la integración de conocimiento en investigación biomédica, con los avances logrados en detección de variantes genéticas y enfermedades asociadas. Aunque varios desarrollos ya cubren algunos aspectos de esta propuesta, consideramos que hay espacio para mejorar las herramientas para las prácticas de salud relacionadas desde un punto de vista de la integración, tanto desde la perspectiva del paciente como de la del médico.

Presentaremos primeramente información de contexto del problema, para dar lugar a una segunda parte donde se incluye el artículo publicado en *Informatics in Medicine Unlocked*. Transcribimos en castellano el *abstract* de dicho artículo:

En este trabajo se presenta un marco para el procesamiento de la literatura genética y genómica, basado en ontologías y recursos léxicos del ámbito biomédico. El objetivo principal es apoyar el proceso de diagnóstico que realizan los médicos genetistas, que extraen el conocimiento de los trabajos publicados. Construimos un oleoducto que reúne varios recursos relacionados con la genética y la genómica y aplica técnicas de procesamiento de lenguaje natural, que incluyen el reconocimiento de entidades nombradas y la extracción de relaciones. Trabajando en un corpus creado a partir de los resúmenes de PubMed, construimos una base de datos de conocimiento que puede ser utilizada para procesar los historias clínicas escritas en español. A partir de una historia clínica de pacientes uruguayos, mostramos cómo podemos mapearla a la base de datos y realizar consultas a nivel grafos para ubicar caminos de conocimiento relevantes. Este framework no es una aplicación de usuario final, sino una estructura de procesamiento extensible que puede ser aprovechada por aplicaciones externas, lo que permite a los desarrolladores de software racionalizar la incorporación del conocimiento extraído.

1. Introducción

El trabajo de un médico especialista en Genética ocurre principalmente en el escritorio, entrevistando pacientes o contrastando conocimientos. Ya sea para prevenir riesgos o diagnosticar enfermedades de causa genética, la lectura de artículos de investigación es una actividad obligada para reunir consideraciones relevantes sobre los últimos hallazgos y evidencias pertinentes a un caso. Está claro que no toda variante genética en el genoma humano es conocida y/o documentada, y clasificar una nueva variante y/o determinar su relevancia clínica, supone un desafío como parte de la evaluación de un caso específico.

Componer un contexto de análisis actualizado con el fin de recomendar los tratamientos, análisis o procedimientos más adecuados demanda una cantidad de tiempo notable que podría suponer varias horas a días, para un paciente en particular.

Idealmente, la historia clínica del paciente debería ser anotada automáticamente aprovechando los últimos conocimientos en todos los campos de la genética, para que el genetista pueda enfocar sus esfuerzos en la elaboración de conclusiones y recomendaciones en base a toda la información disponible.

Cuando se trata de recomendaciones de tratamiento y análisis, la información de punta puede marcar la diferencia para un diagnóstico correcto. Los profesionales de medicina genómica deben tener en cuenta las últimas actualizaciones y hallazgos en el dominio. Las decisiones de los médicos genetistas han de estar respaldadas por las últimos avances en el área, para las cuales hay disponible un número significativo de publicaciones relevantes cada mes.

Se publica una cantidad significativa de artículos vinculados a la Genética humana todos los días. Solo contando PubMed [1], en lo referido a genética, esta cifra es cercana a 300 publicaciones diarias (Figura 1). Esto plantea un desafío para familiarizarse adecuadamente con los últimos hallazgos que podrían ser relevantes para una situación específica del paciente.

Existe un área específica que trabaja con ese tipo de información: Biomedical Text Mining, o BioNLP [2], que se define como el conjunto de métodos de procesamiento de lenguaje natural aplicado a textos de biología médica o molecular. Varios *workshops* y conferencias abordan Procesamiento de Lenguaje Natural para el área de la atención médica, como BioASQ [3], una de las principales competiciones en curso para BioNLP, reuniendo tecnología proveniente del ámbito privado, gubernamental y académico, y con participación de instituciones de la salud (Figura 2).

Este trabajo se concibió teniendo en cuenta la actividad de búsqueda y revisión de literatura útil al médico genetista, quien debe considerar una gran cantidad de hallazgos documentados para recomendar estudios o terapias para cada paciente en particular. A partir de la historia clínica, escrita en español, deberíamos poder acceder a todas las conclusiones relevantes en los documentos resultantes de investigaciones. Las investigaciones producen *abstracts*, y su implicación en una historia clínica particular, se deriva de las entidades y relaciones que de alguna manera enlazan con la misma.

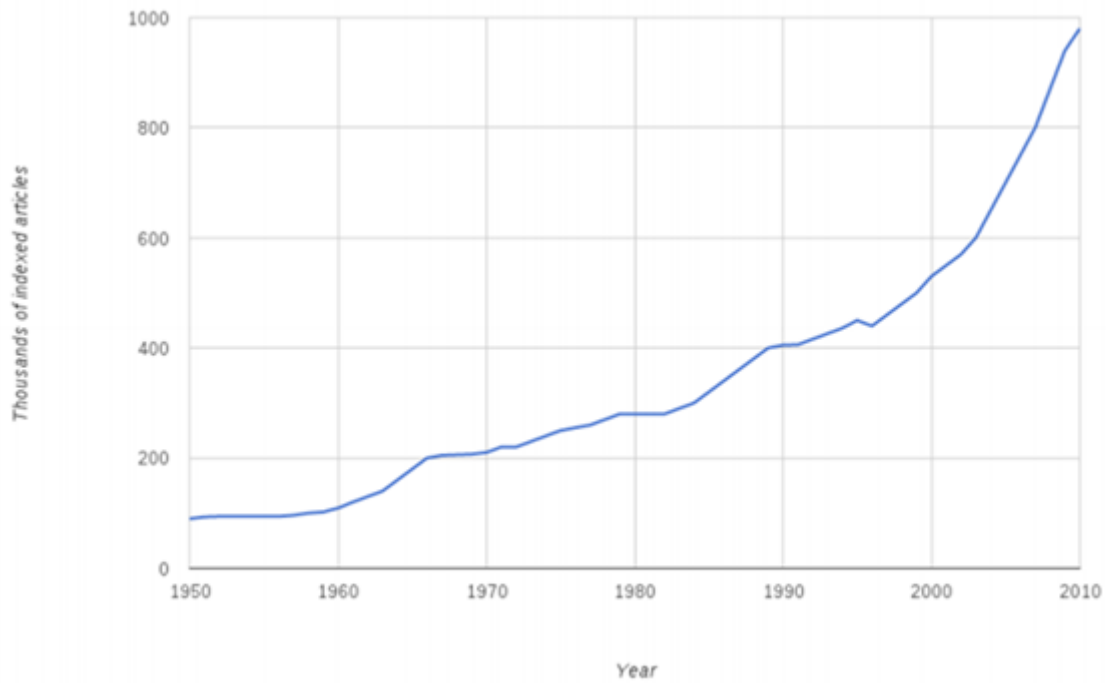


Figura 1 – Evolución histórica de publicaciones de PubMed [4]



Figura 2 - BioASQ – Principales participantes y contribuyentes [3]

El objetivo final de este trabajo es ayudar al médico genetista a tomar decisiones y recomendaciones informadas, al integrar la información basada en entidades y relaciones relevantes, y vincularla a la información contenida en una historia clínica escrita en lenguaje natural. Además, pretendiendo contribuir al desarrollo de estas prácticas en América Latina, nos hemos enfocado en hacer que este motor sea útil para procesar historias clínicas escritas en español, más específicamente, para el léxico médico uruguayo.

Esto nos lleva a pensar en documentos en términos de un modelo de red. Proponemos una integración de documentos, entidades extraídas y relaciones inferidas, en una ontología relativamente simple, que a su vez debe ser sencilla de utilizar desde aplicaciones consumidoras. La Figura 3 ilustra el flujo de trabajo desde el punto de vista funcional.

Nuestra meta entonces puede traducirse en varios objetivos específicos:

- Recopilación de abstracts de PubMed
- Relevamiento e integración de recursos con información sobre genes y fenotipos, en particular fenotipos asociados a enfermedades
- Estudio de enfoques para identificación de entidades y extracción de relaciones a partir de textos científicos
- Análisis de recursos para procesar el texto de historias clínicas escritas en castellano, en particular por médicos uruguayos
- Exposición de los recursos generados por este trabajo, de forma que sean fácilmente utilizables por aplicaciones de software

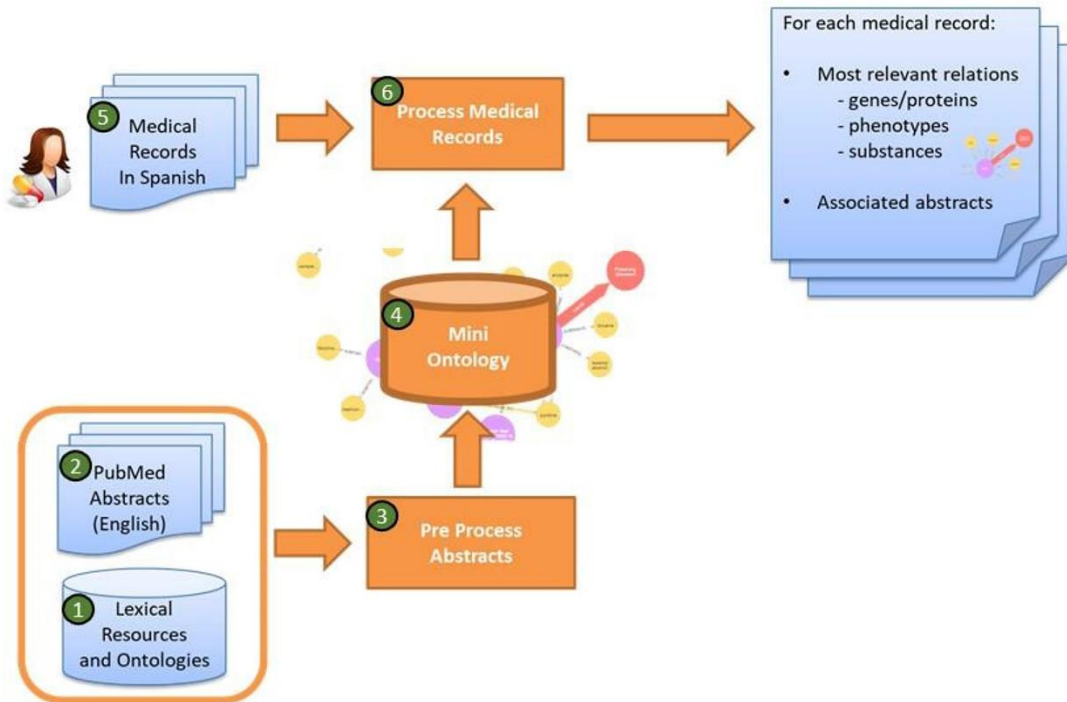


Figura 3 – Marco de trabajo desde la perspectiva del usuario.

2. Recursos y Métodos

Proponemos un marco de trabajo, cuya parte medular está en el reconocimiento de entidades nombradas y construcción de relaciones. Las entidades originalmente son identificadas en un corpus de *abstracts*, para la posterior inferencia de relaciones. Producto de esa fase, se obtiene una base de conocimiento que denominamos *mini ontología*, que será luego utilizada para el análisis de las historias clínicas.

2.1 Recursos

PubMed

Es la fuente para la creación del corpus de abstracts, sobre el que operará el reconocimiento de entidades nombradas, y la extracción de relaciones. Para construir el corpus:

1. Recuperamos resúmenes publicados de PubMed desde 1990 a 2016, que contienen palabras candidatas para nuestro dominio de interés. Los resúmenes, en inglés, que constituirán el corpus de trabajo, se obtienen con una consulta en formato Entrez [5], formulada a partir de los términos:

allele, association, biomarker, chromosome, CNV, complex, congenital, copy-number, diploidy, dna, exome, familial, gene, genetic, genome, genome-wide, genomic, genotype, germline, gwas, hereditary, human, indel, inherited, loci, locus, monogenic, mosaic, mutation, nutrigenetics, nutrigenomics, pharmacogenetics, pharmacogenomics, polymorphism, rare, region, SNP, somatic, spectrum, study, therapy, trait, transcript, variant, variation

2. Se obtiene un corpus que consta de resúmenes correspondientes a 3 millones de documentos, y 582 millones de palabras. Cada resumen suele ser un párrafo relativamente corto (de decenas a unos cientos de palabras), que se pueden dividir en un pequeño número de oraciones. Esto nos lleva a considerar la hipótesis de oración como el alcance general de la inferencia. Por lo tanto, nuestras relaciones extraídas tendrán un alcance dentro de una oración que se almacenará como evidencia para la extracción.

SNOMED CT [6]

Se utiliza con un doble propósito:

- Representación de conceptos base en la mini ontología
- Reconocimiento de entidades nombradas, en español e inglés

SNOMED CT organiza ontológicamente terminología médica, en múltiples lenguajes y localizaciones (Figura 4) por país miembro del consorcio que la mantiene –Uruguay [7] entre de ellos-, siendo hoy la base de terminología médica más extensa a nivel mundial [8]. Existen más de 300000 conceptos, organizados en diferentes clases jerárquicas (Figura 5).



Figura 4 – Versiones localizadas de SNOMED CT. [9]

Hierarchy	Concepts	% of SCT
SNOMED CT Concept (SNOMED RT+CTV3)	321901	100.00%
Body structure (body structure)	31206	9.69%
Clinical finding (finding)	104737	32.53%
Environment or geographical location (environment / location)	1816	0.56%
Event (event)	3614	1.12%
Observable entity (observable entity)	8677	2.69%
Organism (organism)	33696	10.46%
Pharmaceutical / biologic product (product)	17425	5.41%
Physical force (physical force)	170	0.05%
Physical object (physical object)	14841	4.61%
Procedure (procedure)	55880	17.35%
Qualifier value (qualifier value)	9403	2.92%
Record artifact (record artifact)	251	0.07%
SNOMED CT Model Component (metadata)	1568	0.48%
Situation with explicit context (situation)	4277	1.32%
Social context (social concept)	4718	1.46%
Special concept (special concept)	648	0.20%
Specimen (specimen)	1634	0.50%
Staging and scales (staging scale)	1420	0.44%
Substance (substance)	25911	8.04%

Figura 5 – Entidades presentes en la base de datos de SNOMED CT, versión Noviembre 2016 [10]. Se indica para cada tipo de concepto, la cantidad de ocurrencias y qué porción de SNOMED CT representa.

OMIM [11]

Online Mendelian Inheritance in Man (OMIM) es un catálogo de genes humanos, enfermedades genéticas y rasgos fenotípicos (Figura 6), y las relaciones entre ellos, publicado desde 1966, y disponible desde 1995 en la world wide web. Actualmente es mantenido por la universidad Johns Hopkins. Se utiliza con un doble propósito:

- Reconocimiento de entidades nombradas (fenotipos)
- Ground truth para validar relaciones gen-fenotipo

Se encuentra en permanente actualización, constituyendo una referencia de consenso sobre las relaciones entre variantes genéticas y expresión fenotípica. Para cada variante, contiene información de ubicaciones citogenéticas, modos de herencia, y también las mutaciones más relevantes. Puede ser

consultado en línea o a través de una copia de su base de datos actualizada, como ocurre en este trabajo. A la fecha de este trabajo, OMIM contiene unas 26200 entradas, organizadas en los siguientes grupos de loci/fenotipos:

- autosómicos
- ligados al cromosoma X
- ligados en Y
- mitocondriales

*** 605882**

BRCA1-INTERACTING PROTEIN 1; BRIP1

Alternative titles; symbols

BRCA1-ASSOCIATED C-TERMINAL HELICASE 1; BACH1
 DELETIONS OF GUANINE-RICH DNA, C. ELEGANS, HOMOLOG OF
 DOG1, HOMOLOG OF
 FANCI GENE; FANCI

HGNC Approved Gene Symbol: BRIP1

Cytogenetic location: 17q23.2 Genomic coordinates (GRCh38):
17:61,679,185-61,864,119 (from NCBI)

Gene-Phenotype Relationships [View clinical synopses as a table](#)

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
17q23.2	Fanconi anemia, complementation group J	609054		3
	{Breast cancer, early-onset, susceptibility to}	114480	SMu, AD	3

Figura 6 – Ejemplo de entrada en OMIM: gen BRCA1, y fenotipos relacionados. <https://www.omim.org/entry/605882>

Varnomen/HUGO/HGVS [9]

La organización Human Genome Organisation (HUGO) fomenta la colaboración científica, en el contexto del Proyecto Genoma Humano. Human Genome Variant Society (HGVS) impulsa la caracterización y documentación de variantes humanas. Ambas organizaciones promueven el estándar internacional de nomenclatura de variantes Varnomen, aplicable a ADN, ARN y proteínas.

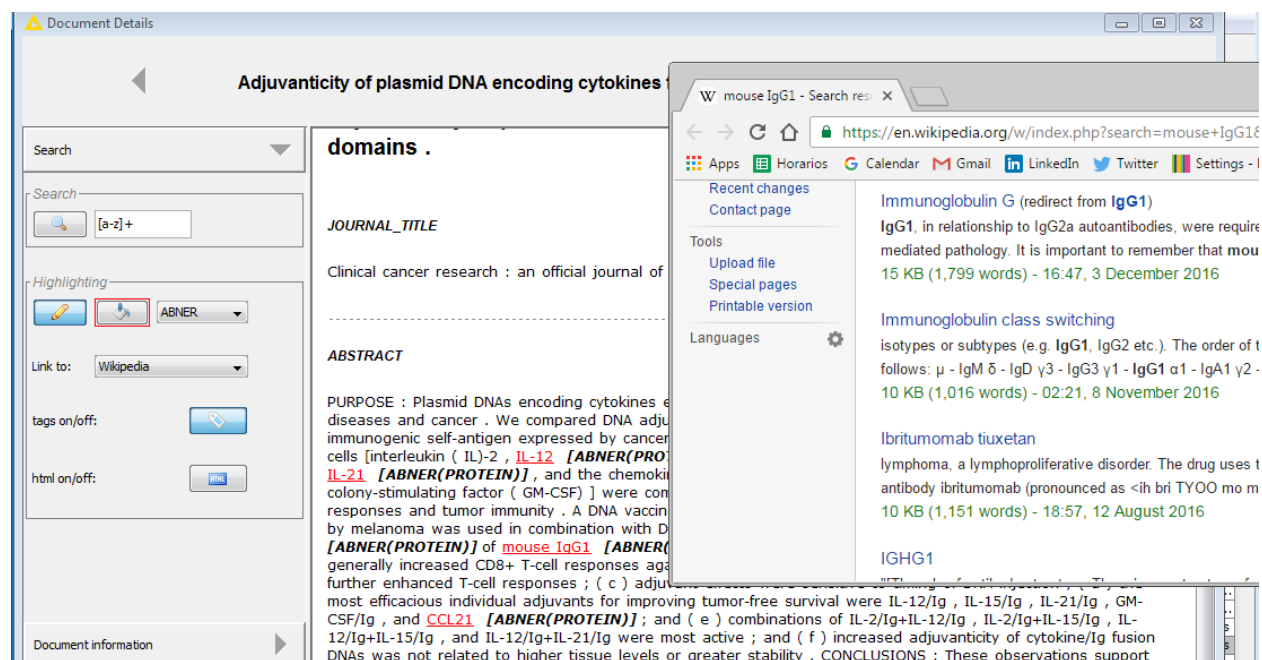
Resúmenes de Historias Clínicas, en español de Uruguay

Para este trabajo, no fue posible conseguir historias clínicas digitales, por lo que exploramos fuentes alternativas. Tomamos entonces como referencia de texto de historias clínicas, los resúmenes disponibles en los Encares publicados por la Oficina del Libro de Facultad de Medicina de la Universidad de la República:

- Encares de Clínica Neurológica 1 [12]
- Encares de Clínica Neurológica 2 [13]
- Encares para el Internado Obligatorio [14]
- Hematología [15]
- Residencia en medicina interna [16]
- Encares de Clínica Cardiológica [17]

ABNER tagger [5]

ABNER es una herramienta para reconocimiento de entidades nombradas (genes y proteínas) (Figura 7), basado en *conditional random fields*, y entrenado a partir del corpus Biocreative [19].



The screenshot displays the ABNER tagger interface. The main window shows a document titled "Adjuvanticity of plasmid DNA encoding cytokines" with several entities highlighted in red boxes: "mouse IgG1", "IL-12", "IL-21", and "CCL21". The interface includes a search bar, a highlighting tool, and a link to Wikipedia. A browser window in the background shows search results for "mouse IgG1" on Wikipedia, with the first result being "Immunoglobulin G (redirect from IgG1)".

Figura 7 – Ejemplo de entidades identificadas por Abner en el texto

2.2 Métodos

2.2.1 Procesamiento de Abstracts

Para la identificación de entidades y extracción de relaciones en abstracts, el corpus de trabajo está basado en PubMed según se detalló en la sección anterior.

Entidades

Los procesos de identificación de entidades utilizan diccionarios elaborados a partir de diversos tesauros disponibles:

1. Enfermedades y sustancias

Utilizamos diccionarios obtenidos a partir de SNOMED CT en sus ediciones Internacional (en inglés) y Uruguay, para identificar enfermedades y sustancias.

2. Genes y proteínas.

Utilizamos un diccionario generado a partir de la nomenclatura de genes Varnomen/HUGO/HGVS, además de ABNER Tagger. Asumimos genes y proteínas como un mismo tipo de entidad, dado que a menudo una proteína (como expresión funcional de un gen) toma el nombre de ese gen; aún en los casos en que no es así, hay una relación uno a uno entre el nombre (o símbolo) de un gen y los de una proteína.

3. Fenotipos

Utilizamos un diccionario para nomenclatura de fenotipos en humanos, obtenido a partir de OMIM.

Relaciones

La estrategia para extracción de relaciones entre las entidades se compone de varias tareas de procesamiento de lenguaje natural:

1. Tokenizing – separación en oraciones y palabras
2. Stemming – raíz morfológica de las palabras
3. Las relaciones a buscar están expresadas con reglas (Figura 8), donde se tiene:
 - Tipo de relación – por ejemplo, *causa*
 - Tipo de las entidades relacionadas – por ejemplo, *enfermedad y proteína*
 - Palabra “señal” – por ejemplo, *causante*

Para mejorar la precisión de la identificación de las relaciones, utilizamos un clasificador binario, basado en la idea de Supervisión Distante [20]. Para ello:

1. Se toma como *ground truth* las relaciones explícitamente indicadas en OMIM MorbidMap [21]
2. Se generan atributos en el texto:
 - *Part of Speech* en una ventana de palabras (+/- 2 desde la entidad identificada)
 - Distancia
 - Lema
3. Se prueba con diferentes modelos de aprendizaje automático (KNN, SVM, MLP). Esto se detallará en un capítulo específico.

S TipoRel	S TNE1	S Cue	S TNE2
ASOC	PROTEIN GENE-SYMBOL	associated	DISEASE
CAUSA	PROTEIN GENE-SYMBOL	caused caused by mutations causing underlies linked directs	DISEASE
CAUSA	GENE-SYMBOL PROTEIN	cause caused lead to leading to responsible involved	DISEASE
CAUSA	GENE-SYMBOL PROTEIN	mutated underlies family Presenting dysfunction effect	DISEASE
INFSUS	SUBSTANCE	effect impact related associated with sensitivity determined influence weight genetic ...	GENE-SY...
INFSUS	SUBSTANCE	response role link between association moderation relationship between response	GENE-SY...
SURI	CHROM	susceptibility	DISEASE
SURI	PROTEIN GENE-SYMBOL	associated Association risk contributes susceptibility	DISEASE ...
INFSUS	PROTEIN GENE-SYMBOL	dose acquire express alter	PHENOTY...
INFSUS	PROTEIN GENE-SYMBOL	dose acquire express alter	DISEASE

Figura 8 – Reglas para identificación primaria de relaciones

2.2.2 Procesamiento de Resúmenes de Historias Clínicas

Para los resúmenes de historias clínicas, se identifican las entidades nombradas, típicamente enfermedades. Para esto, se utiliza un subconjunto de las herramientas y técnicas citadas anteriormente, contando con diccionarios generados a partir de SNOMED CT versión Uruguay.

2.3 Notas de Implementación

La base de conocimientos se inicializa a partir de los conceptos existentes en SNOMED CT. Estos están expresados como conceptos base (FSN, en SNOMED), y disponibles en formato base de datos relacional.

Cada concepto se vincula con sus nombres alternativos (sinónimos), incluyendo denominaciones en diferentes idiomas. Adicionalmente, se incorporan las relaciones existentes entre conceptos (ejemplo: Figura 9).



Figura 9 - SNOMED CT: algunos sinónimos y conceptos relacionados con Aspirina (producto)

Esta inicialización a partir de SNOMED CT incorpora unos 3.9 millones de relaciones (Tabla 1).

Relación	Ocurrencias	%
Is_a	1.588.352	40,9%
Finding_site	360.071	9,3%
Associated_morphology	333.701	8,6%
Morphology	333.701	8,6%
Method	187.914	4,8%
Part_of	166.654	4,3%
Procedure_site_-_Direct	133.341	3,4%
Interprets	90.778	2,3%
Causative_agent	77.657	2,0%
Has_active_ingredient	52.652	1,4%
Component	42.552	1,1%
Occurrence	40.436	1,0%
Procedure_site_-_Indirect	33.855	0,9%
Finding_method	30.827	0,8%
Direct_morphology	24.734	0,6%
Has_definitional_manifestation	24.504	0,6%
Has_dose_form	21.233	0,5%
Pathological_process	18.860	0,5%
Laterality	18.400	0,5%
Using_device	17.367	0,4%
Otras	289.339	7,4%
Total	3.886.928	100,0%

Tabla 1 - Ranking de ocurrencias de relaciones en SNOMED CT, y su peso relativo en el total.

Para la construcción de los mencionados procesos de se utiliza:

- KNIME [22] – plataforma de integración de algoritmos de data science
- Neo4J [23] – base de datos de grafos
- Java
- Freeling [24] – herramienta para tareas de PLN

3. Artículo

El artículo está organizado en 6 secciones:

1. Introducción
Se presenta la motivación y objetivos del trabajo.
2. Antecedentes
Revisión de publicaciones y recursos relevantes.
3. Materiales y métodos
Estrategia, herramientas y recursos léxicos que se utilizaron en la implementación.
4. Resultados
Se incluyen algunas métricas sobre partes del proceso.
5. Discusión y conclusiones
Comentarios sobre beneficios, limitaciones y notas de uso.
6. Trabajo futuro
Posibles próximas acciones de profundización a partir de este trabajo.

[Vínculo al artículo](#)



Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

From medical records to research papers: A literature analysis pipeline for supporting medical genomic diagnosis processes[☆]



Fernando López Bello^{a,*}, Hugo Naya^b, Víctor Raggio^c, Aiala Rosá^d

^a Basic Sciences Development Program (PEDECIBA, www.pedeciba.edu.uy), ICT4V (ict4v.org), Av. Italia 6201, Parque Tecnológico del LATU, Montevideo, 11500, Uruguay

^b Bioinformatics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay

^c Genetics Department, Facultad de Medicina, Universidad de la República de Uruguay, Montevideo, Uruguay

^d Computer Science Institute (Instituto de Computación, INCO), Engineering School (Facultad de Ingeniería, FING), Universidad de la República, Montevideo, Uruguay

ARTICLE INFO

Keywords:

Controlled vocabulary
Natural language processing
Genomics
Automated pattern recognition
Publications
Medical records

ABSTRACT

In this paper, we introduce a framework for processing genetics and genomics literature, based on ontologies and lexical resources from the biomedical domain. The main objective is to support the diagnosis process that is done by medical geneticists who extract knowledge from published works. We constructed a pipeline that gathers several genetics- and genomics-related resources and applies natural language processing techniques, which include named entity recognition and relation extraction. Working on a corpus created from PubMed abstracts, we built a knowledge database that can be used for processing medical records written in Spanish. Given a medical record from Uruguayan healthcare patients, we show how we can map it to the database and perform graph queries for relevant knowledge paths. The framework is not an end user application, but an extensible processing structure to be leveraged by external applications, enabling software developers to streamline incorporation of the extracted knowledge.

1. Introduction

1.1. Motivation

With the advent of next-generation sequencing technologies, a vast amount of genomic data is generated for each patient. It is well-known that this trend surpasses Moore's Law in terms of cost, therefore facilitating the incorporation of genomic sequencing into routine clinical studies. A patient genome can be sequenced in a few hours or even minutes [1], and then is annotated using available genetic databases. Nevertheless, there is still further work needed before the physician has all of the necessary elements readily available for a specific case. The literature requires reviewing in such a way that will allow the gathering of the latest findings, including gene, gene expression, and gene-disease associations.

The diagnosis process of medical geneticists must consider traits and incident factors, which starts from the inherited alleles downstream of the observed phenotypes, which are frequently noticed as diseases or

health conditions. Generally speaking, genes cause phenotypic variation, which is not necessarily pathogenic. The clinical expression of a gene variant, its interaction with non-genetic factors and substances, and its causative relationship with disorders or health conditions are progressively being documented as a result of ongoing research. From a medical standpoint, it is important that a medical geneticist can seize the specific knowledge that applies to an individual and his/her genome. Typical starting questions are: *Is this variant pathogenic?*; *With which phenotypes/diseases is this variant associated?*; *Are there known substance/drug interactions?*; and *How is it associated with other traits or medically relevant conditions or risks?*

A vast number of relevant publications seed a constantly growing knowledge set relevant to these medical genomics-related areas. Only counting PubMed, this figure is close to 300¹ publications per day. This makes attempts to become properly acquainted with the latest findings that could be relevant to a specific patient particularly challenging.

[☆] Thanks to Dr. Laura Rodríguez (MD), who collaborated with an exhaustive analysis of the process outputs for evaluating this work's results.

* Corresponding author.

E-mail address: fernando.lopez@ict4v.org (F. López Bello).

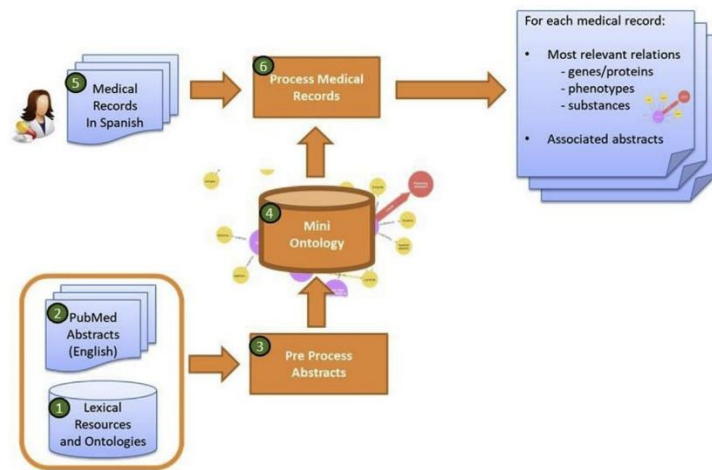
¹ This figure corresponds to the number of abstracts retrieved from PubMed [2] for Genetics, with the query shown in Fig. 4 – Document Grabber Process, on a daily basis.

<https://doi.org/10.1016/j.imu.2019.100181>

Received 19 February 2019; Received in revised form 26 March 2019; Accepted 7 April 2019

Available online 13 April 2019

2352-9148/© 2019 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



document collection. Access APIs: Neo4J API (REST, Python, Java), Solr API (REST, Java)

5. Medical Records are received in an inbox, currently in CSV format. Tools used: KNIME

6. Process Medical Records block identifies entities in Medical Records, and maps them to the existing Mini Ontology entities. Tools used: KNIME, Neo4J, Solr

1.2. Objective

The goal of this work is to provide tools for the medical geneticist that optimize his/her access to the latest research pertaining to a specific patient (or to specific genomic information). Given some clinical and genomic information, the medical geneticist's diagnostic process should benefit from a toolset that eases discovering useful knowledge, such as related disorders, genetic causes or predispositions, associated proteins or substance interactions. A typical usage sequence is depicted in Fig. 1.

That said, the main product of our work is a framework for integrating domain sources and extracting useful knowledge from them, proposing a workflow for extending it with new sources. Furthermore, it should be possible to build applications that leverage this product with open application programming interfaces (APIs).

2. Background

2.1. Corpus of research abstracts

For our work, PubMed [2] was the main corpus of reference, with a yearly publication rate close to one million articles, being used for the preparation of more specific annotated benchmarks and corpora. Currently, PubMed features over 25 million XML formatted abstracts, including all MEDLINE abstracts (journals files since 1946) [3].

Since these documents contain text, written in natural language, to extract information from them requires Natural Language Processing (NLP) techniques. Several workshops and conferences address NLP for biomedical texts – the so-called BioNLP. A good example focusing on healthcare is BioASQ [4], one of the main ongoing competitions on the subject, gathering private, government, and academic technology and health-related institutions.

2.2. Challenges and resources

A usual challenge in BioNLP is the great variety and ambiguity of the terminology. Here is where lexical resources, such as thesauri and ontologies, specific to this domain, become a fundamental tool for the whole process.

In the healthcare domain, terminology standards have achieved

notable maturity. Remarkable cases include ICD [5] for term classification, and SNOMED CT [6] as a multi-language collection of ontology organized hierarchies, also custom-built and extended for various affiliated countries.

More specifically, SNOMED CT is a collection of medical term thesauri, with an ontology structure—including synonyms, hyponyms, and hypernyms for each term. SNOMED CT comprises over 400 K concepts and 7.5M relations, which can be mapped to preexisting standards, such as ICD-9. There is an adapted country version for Uruguay [7].

Gene and protein nomenclature normalization efforts have produced well-defined standards, such as VarNomen/HUGO/HGVS [8] and UniProt [9]. Nevertheless, BioNLP papers quite often show different denominations for the same gene or protein. For example, the *Homo sapiens* gene WASHC5 [10] (also known as KIAA0196) codes for protein WASH-complex-subunit 5 (also known as Strumpellin).

Abbreviations pose an additional challenge in BioNLP, with its disambiguation being a context-sensitive problem. Thus, for example, the symbol T can refer to both T-cells and the T-gene. Liu et al. [11] obtain an abbreviation accuracy of 82% on this matter.

2.3. Entity recognition works in biomedical texts

An important reference for this work addresses the extraction of relevant entities and potential relations, targeting the identification of risk factors and associated pathologies [12]. The cited article explores the power of combining tools and sources, such as MeSH, Genia Tagger, SNOMED CT, and MEDLINE, thus acting as a capstone for the idea that we develop in this work. As we will see, several of these components play a role in our processing pipeline. Genia Tagger [13] has been frequently used both for part of speech tagging and named entity recognition (NER) in the BioNLP domain. For Spanish medical documents, Genia Tagger has been used in conjunction with Freeling [14] for entity recognition and automatic annotation [15]. ABNER [16] is another often-used gene and protein tagger with good performance, based on Conditional Random Fields. Additionally, several approaches rely strongly on gazetteers for precise entity identification (e.g., Hina et al. [17]).

Some efforts have addressed automatic or semi-automatic annotation of medical records. For the NER task, several mixed approaches combine Conditional Random Fields (CRF) and thesauri or ontologies

Fig. 1. Proposed framework usage from a user perspective. The user input consists of medical records, written in Spanish. Using a pre-generated knowledge database, each medical record is linked to the most significant relations extracted from research articles. As a result, on output, the framework produces a graph that integrates the main medical concepts of the medical record and applicable relations inferred from abstracts. The main steps are:

1. PubMed Abstracts (English) are acquired from PubMed by a process that downloads, unpacks and parses documents. Tools used: KNIME, Entrez
2. Lexical Resources and Ontologies were previously downloaded. Resources used: SNOMED CT, OMIM, HGVS
3. These two resources are used by the Pre Process Abstracts block, which builds the Mini Ontology graph with recognized entities and relations. At the same time, it builds a document collection form information retrieval. Tools used: KNIME, Abner Tagger, Java, Neo4J, Solr.
4. The Mini Ontology is available as graph and

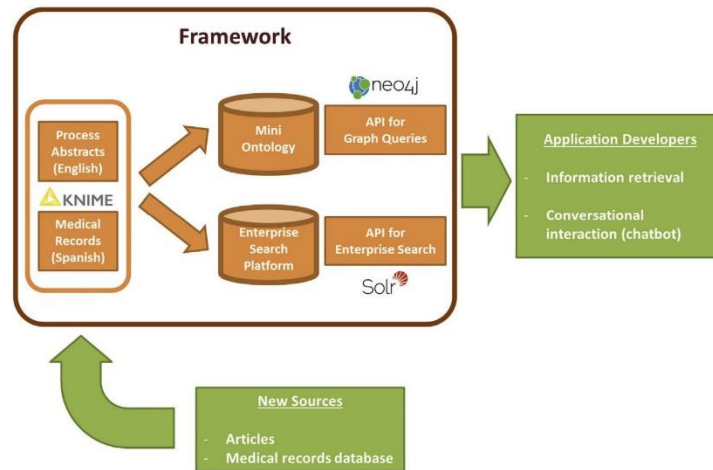


Fig. 2. Framework context: Ingestion from text sources and application development.

[18,19]. In addition to those methods, others (e.g., Chikka et al. [20]) achieved a notable performance in information extraction, making use of specifically trained text mining models (e.g., using Apache cTAKES [21]).

Wu et al. [22] compare alternatives based on word embeddings to improve NER results in BioNLP, against existent proposals based on CRF, MaxEnt, and SVM. Chiu et al. [23] devise guidelines for good word2vec based embeddings, both CBOW and skip-gram, working on PubMed and the PMC corpus. For auxiliary tasks, these authors use GeniaSS as a sentence splitter and NLTK [24] for word tokenizing.

Gong [25] discusses the current application of text mining alternatives to the biomedical domain, proposing a pipeline for recognizing biomedical concepts, as well as discussing core components, such as POS tagging and CRF-based (like ABNER) or support vector machines, and how to combine machine learning with the dictionary-based approach. Several of these processing concepts are present in our work.

2.4. Relation extraction

Once the problem of identifying named entities has been addressed, knowledge can be expressed in terms of relations within each pair of entities, conforming tuples:

{relation_type, entity_1, entity_2}.

Among the vast number of publications regarding relation extraction, it is worth mentioning a few examples to contextualize our work. Open information extraction initiatives, such as Never Ending Language Learning (NELL) [26] or question answering targeted IBM Watson [27], present broad scope solutions for relation extraction, where relationship types might not be completely known in advance. In our case, we are focusing on a more specific context, with the defined entity and relationship types (e.g., genes and substances) and the interactions (e.g., cause or association) between them. An example of such a scenario can be found in BeFree [28]. Besides this, we also considered distant supervision. According to Mintz et al. [29]: 'The intuition of distant supervision is that any sentence that contains a pair of entities that participate in a known [...] relation is likely to express that relationship in some way'. It is a relation learning strategy that builds a machine learning model from known ground truth, with pre-annotated relations, to classify a relation as existent or nonexistent. This model evaluates multiple features extracted from the natural language sentence.

2.5. Integration

On knowledge integration, NIH/NLM's SemRep [30] includes a relation extraction tool, from which the SemMedDB [31] database is generated. SemMedDB is a set of relations extracted from MEDLINE and includes a visualization browser, providing a concrete example of what we expect from a knowledge ontology. However, it is worth noting that our approach aims to integrate disparate sources, incorporating lexicon specific to genetics and genomics, and regionalized resources, such as SNOMED CT per-country releases.

The extracted information is traditionally delivered through a relational database, a dedicated framework [32] for querying normalized text or a data warehouse infrastructure [18]. For our proposal, we considered existing ontology graph representations for healthcare scenarios [33,34].

3. Material and methods

3.1. General framework description

Recapping Fig. 1, we built a framework capable of processing medical records written in Spanish, and producing an output that synthesizes literature findings that are relevant to it, from a genetics perspective. This knowledge synthesis is composed of concepts (entities) and relevant relations between them, as well as the actual document reference that backs each of these results. Actually, this output is the subset of a major knowledge database that has been preprocessed in advance.

Entities can be genes or proteins, phenotypes, substances or disorders. The main knowledge base, called Mini Ontology, is permanently updated from a corpus that contains novel articles, on a daily basis. Each new document from the corpus can eventually produce new relations between entities. Currently, the corpus consists of genomics-related PubMed abstracts (about 3 million abstracts in English, from 1990 until 2018), but can be extended to new sources. Also, each new medical record can be integrated into the knowledge base for further exploration.

The knowledge base can be queried in two ways: 1) starting from the medical record, and leading to related entities (like genes); or 2) starting from genes of interest (previously obtained from patients genome or exome analysis), and leading to related diseases and

substances.

The framework exposes services that application developers can leverage, and can easily integrate new data sources (Fig. 2).

As the main resource, the framework builds a knowledge base (Mini Ontology) from article abstracts and can be used to either traverse genomics concepts and articles in an ontology fashion or assess the medical genetics literature about relevant findings related to a specific medical record, written in Spanish.

The Mini Ontology can be accessed from external applications via Application Program Interfaces (APIs) that include graph queries and enterprise search requests, from almost every modern programming language. Before providing the pipeline details, it is worth noting that there are several integration points that can be used to bring new sources on board: 1) Abstract articles are currently received from PubMed, but the same inbox can receive documents from other sources. For extraction from PubMed, we built a custom process that could eventually fetch literature from other sites or news feeds. 2) Tools for NER currently include several statically bound dictionaries, but this set can easily be extended with new ones, to handle more terminology. 3) The Mini Ontology relies on the ontological definition of the *concept* from SNOMED CT, which is language independent.

3.2. Implementation

Our proposal relies on a combination of curated thesauri and dictionaries, NLP techniques, and a graph database engine. At its core, building the Mini Ontology includes these main components (see Fig. 3): Abstracts Processing (NER and Relation Extraction, in English); Medical Records (NER, in Spanish), and Mini Ontology Integration.

The development platform included the following tools: KNIME [35], a Java-based open source data science platform; Neo4j graph engine [36] with Cypher query language.

Apart from the present work, additional Enterprise Search

functionality was implemented with Apache Solr [37], which builds an index on top of the Mini Ontology to enable Solr JSON queries.

3.2.1. Obtaining PubMed documents

Our working corpus consists of about 3 million PubMed abstracts, which are retrieved from PubMed on a periodic basis (e.g., daily) by submitting a query against the PubMed database, using the Entrez search and retrieval system [38]. The query defines a document subset specific to Human Genetics.

Our document retrieval process is implemented in KNIME (Fig. 4, Document Grabber Process), and produces Comma Separated Value (CSV) files containing the abstracts, tokenized with OpenNLP Word Tokenizer [39]. This CSV is transformed into an internal representation format containing: Article ID, Title, Abstract Text, Authors, Source, and Release Date.

We take into account that other sources, beyond PubMed, could be added into our pipeline. In any case, the new source must be adapted to the above-defined format, which can be easily achieved by leveraging existing KNIME functionality, such as PDF-to-text conversion based on Tika [40], RSS feeds, and REST requests.

3.2.2. English Abstracts Processing

Typically, our corpus abstracts are relatively short paragraphs, from tens to a few hundreds of words. Each relation is scoped within a sentence. This stage is also implemented in KNIME, integrating NER and relation extraction strategies as described below.

3.2.2.1. *Named entity recognition (NER)*. We rely on SNOMED CT ontologies to identify entities, such as disorder, finding, substance, product, organism, and morphological abnormalities. Acronym and pre-coordination disambiguation tasks are part of this workflow.

We also used OMIM [41], a curated database of gene-phenotype relations, which includes gene name homonyms. It is worth noting that

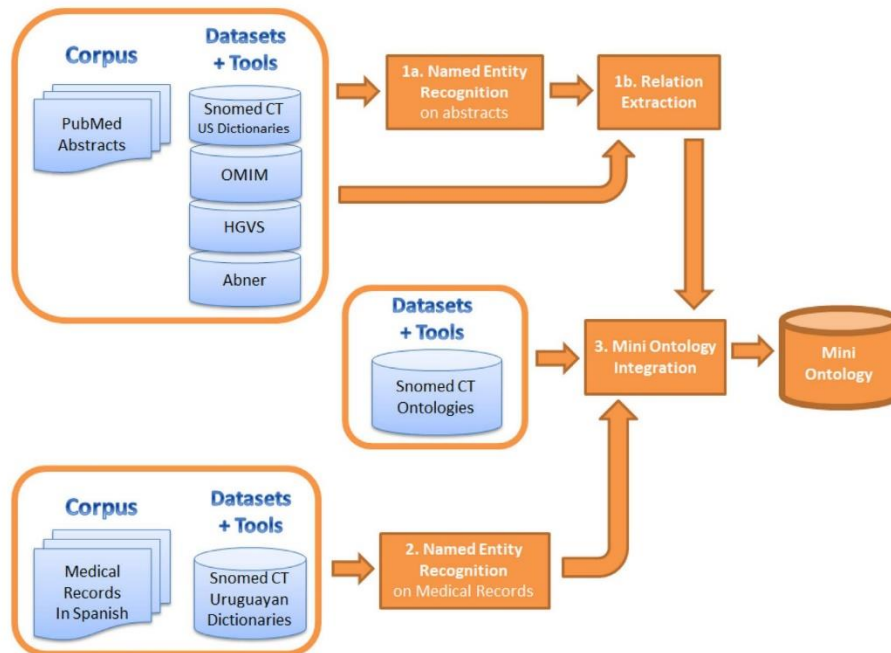


Fig. 3. Process pipeline for generating the Mini Ontology from corpora and datasets.

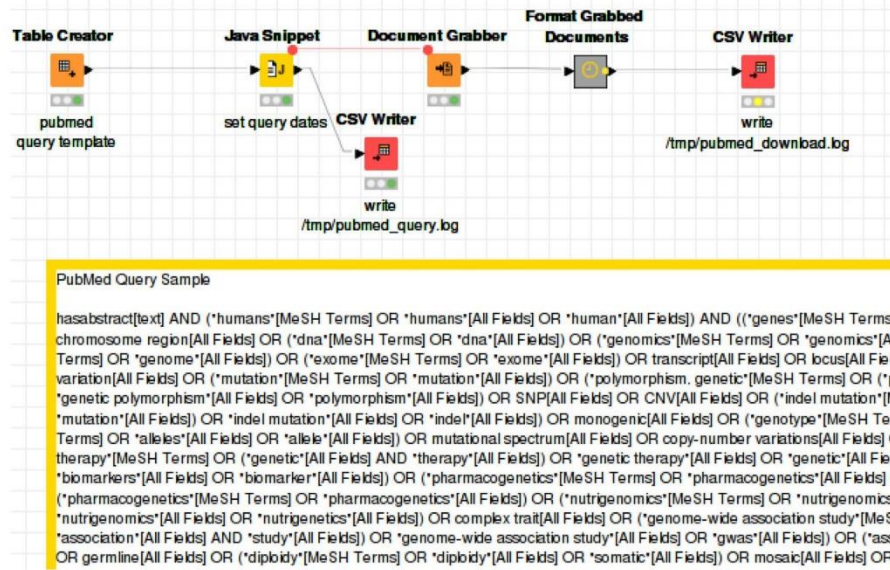


Fig. 4. Document Grabber process Excerpt.

OMIM enumerates specific phenotypes (which might be a disorder or not) related with genetics, and thus—for our specific case—is a more precise source for phenotype identification than SNOMED CT.

Finally, VarNomen-HUGO-HGVS provides a standardized symbol nomenclature for genes and proteins. Once the abstracts corpus is ready, initial tasks include sentence splitting, case conversion, and special characters removal, translating foreign symbols. A cascade of dictionary-based identification is applied for initial entity identification. Then, we use ABNER tagger to recognize genes and proteins. ABNER tagger is based on Conditional Random Fields, and in our case, it is configured to use the BioCreative corpus model.

3.2.2.2. *Relation extraction.* For relation extraction within our pipeline, we propose a simple heuristics-based mechanism that leverages available domain-specific resources and also lets domain experts establish basic rule sets for identifying significant relations.

Rules for inferring relation are specified by indicating two *entity types* and one or more *cue words*, as shown in Table 1. To improve recall, cue words are reduced to its word stems. This way, a user can specify as many plain rules as needed, later expanded into single, atomic rules in the form < Relation Type, Entity 1 Type, Cue Word Stem, Entity 2

Type > .

We specified the following relation rules:

- **Cause:** An entity (*protein, gene, substance or phenotype*) that causes a *disease*
- **Association:** A generic association between an entity (*protein, gene, substance or phenotype*) and a *disease*, not necessarily implying causative effects
- **Substance influence:** A *substance* that is relevant for functional aspects of a *gene or protein*
- **Susceptibility or risk:** A *protein or gene* that poses a risk of triggering a *disease or phenotype*

For example, for “Heparin-induced thrombocytopenia is caused by the formation of antibodies that bind to specific complexes of platelet factor 4 (PF4) and heparin”, two cause (CAUSE) relations are inferred, as per the second rule defined in Table 1:

CAUSE defined by: gene-symbol = pf4 (PF4) / disease = heparin induc thrombocytopenia (Heparin induced thrombocytopenia) / cue = caus by

Table 1
Rules for detecting relations.

Rel.Type	Entity 1 Type	Cue Words	Entity 2 Type
ASOC	PROTEIN GENE-SYMBOL	associated	DISEASE
CAUSA	PROTEIN GENE-SYMBOL PHENOTYPE	caused caused by mutations causing underlies linked directs	DISEASE
CAUSA	GENE-SYMBOL PROTEIN	cause caused lead to leading to responsible involved	DISEASE
CAUSA	GENE-SYMBOL PROTEIN	mutated underlies family Presenting dysfunction effect	DISEASE
INFSUS	SUBSTANCE	effect impact related associated with sensitivity determined influence weight genetic marker responsive loci predicting outcome relationship efficacy based on	GENE-SYMBOL PROTEIN
INFSUS	SUBSTANCE	response role link between association moderation relationship between response	GENE-SYMBOL PROTEIN
SURI	CHROM	susceptibility	DISEASE
SURI	PROTEIN GENE-SYMBOL	associated Association risk contributes susceptibility	DISEASE PHENOTYPE
INFSUS	PROTEIN GENE-SYMBOL	dose acquire express alter	PHENOTYPE
INFSUS	PROTEIN GENE-SYMBOL	dose acquire express alter	DISEASE

CAUSE defined by: protein = platelet factor 4 (platelet factor 4) / disease = heparin induc thrombocytopenia (Heparin induced thrombocytopenia) / cue = caus by}

Finally, to improve precision, we refined the relation extraction strategy using distant supervision, as explained in section 4.3.1 Improving Precision.

3.2.3. Medical records: NER, in Spanish

The expected input of the framework, either from a user or an external application, are medical records written in Spanish. We perform an NER recognition on them, to map Medical Records with the existing knowledge base. This step is implemented in KNIME.

Since an actual set of medical records in Spanish was not available for research, we manually transcribed 109 clinical notes with physician observations, from actual patient cases, used for medical education. Original stories come from university course textbooks from the Faculty of Medicine (Universidad de la República, Uruguay), covering multiple topics: Hematology (Hematología), Neurology (Encares de Clínica Neurológica 1, Encares de Clínica Neurológica 2), Internal Medicine (Encares para el Internado Obligatorio, Residencia en medicina interna), and Cardiology (Encares de Clínica Cardiológica)

As a government healthcare agency (Salud.Uy) initiative, a full-fledged SNOMED CT release is available for Uruguayan Spanish terminology, for translating existing terms, and constantly incorporating new ones. This gives us the chance to elaborate specific dictionaries for entity recognition, which are specific to medical records spellings: Findings, Substances and Product Names, Organisms, Morphologic Abnormalities, and Disorders.

Most of these disease term recognition tasks are performed based on the dictionary approach. Additionally, we attempted an additional step for further recognition using KNIME's nodes for Stanford Named Entity Learner [42], for which we trained a CoreNLP CRF model, but had issues recognizing multi-term diseases. For example, for *hipertensión arterial* (English *arterial hypertension*), the dictionary-based approach correctly recognizes *hipertensión arterial* disease in Spanish, but the CoreNLP node recognized both *hipertensión* (English *hypertension*) and *arterial* only individually.

3.3. Mini Ontology Integration

Relevant relations, medical records, and PubMed abstracts corpora are integrated into the depicted graph structure (Fig. 5).

The relation triplets produced so far carry their source information as attributes to enrich the graph of our ontology: publication ID, authors, source, abstract title, and a sentence from which the relation was inferred.

All of this information is exported to the graph database, which was implemented in the Neo4J graph engine. The graph can be queried and visualized for inquiry and exploration. At this point, querying from a medical record vertex to a PubMed entity by using the inferred relations as the valid path should convey the expected end-to-end result, as depicted in Fig. 6. The graph can be traversed from the medical record to PubMed entities, through inferred relations. In this example:

- A medical record concept (gray node with ID 7010), has a relation that points to the Spanish SNOMED CT translation *glomerulonephritis (trastorno)* (green node). The medical record attributes include the source sentence.
- Node *glomerulonephritis (trastorno)* is a translation from the main SNOMED CT concept *Glomerulonephritis (disorder)* (green node, bigger size).
- *Glomerulonephritis (disorder)* SNOMED CT concept is a reference from the *Glomerulonephritis* disease entity (red node), found in a PubMed abstract.
- Another PubMed entity, *IgA*, is a protein/gene (violet node) related

(CAUSE relation) with the *Glomerulonephritis* disease entity. The CAUSE relation attributes include the source sentence and PubMed article ID.

In summary, from the *medical record nodes*, we reach a *causative relation* that involves a *gene/protein node*.

4. Results

4.1. Mini ontology dimensions

To give an idea of the size of the built knowledge base, here we show a few count measures of the graph.

- **Entities.** Nodes correspond to entities that are either preloaded from SNOMED CT or recognized at PubMed abstracts or medical records.

NER Evaluation Summary – Asthma	
Retrieved Entities	4.147
True Positive	2.498
False Positive	1.649
Precision:	60%

- **Relations.** Edges represent relations, either pre-loaded from SNOMED CT or inferred from PubMed.

Correct or nearly correct entities, counting repetitions:	
Identified named entities:	812
Relevant entities:	760
Precision:	94%

Strict criteria (correct or nearly correct entities, not counting repetitions):	
Identified named entities:	403
Relevant entities:	351
Precision:	87%

4.2. NER evaluation

4.2.1. PubMed abstracts

To evaluate NER, we selected *asthma*, a common disorder with potential genetics causes or associations, and which is a disorder family.

We focus on the named entities that take part in relations involving genes or proteins. For the sake of better coverage, we consider relaxed criteria, in which we accept gene-related concepts as valid entities (for example: “5 lipoxygenase promoter”, “A1A2 genotype”, “A2 genotype”).

This approach produced a rather fair result for *precision* (see Table 2). On the other hand, calculation of the *NER recall* requires manually identifying entities in actual abstracts corpus that were not automatically recognized. We consider this additional effort as beyond the scope of this work.

4.2.2. Medical records, in Spanish

The simple NER approach that we applied to medical records, leveraging Uruguayan terminology dictionaries from SNOMED CT, produced the results listed in Table 3 for the domains of *disease* and *finding* entities. A physician manually evaluated the expected vs. found entities for each medical record, considering an entity as correctly

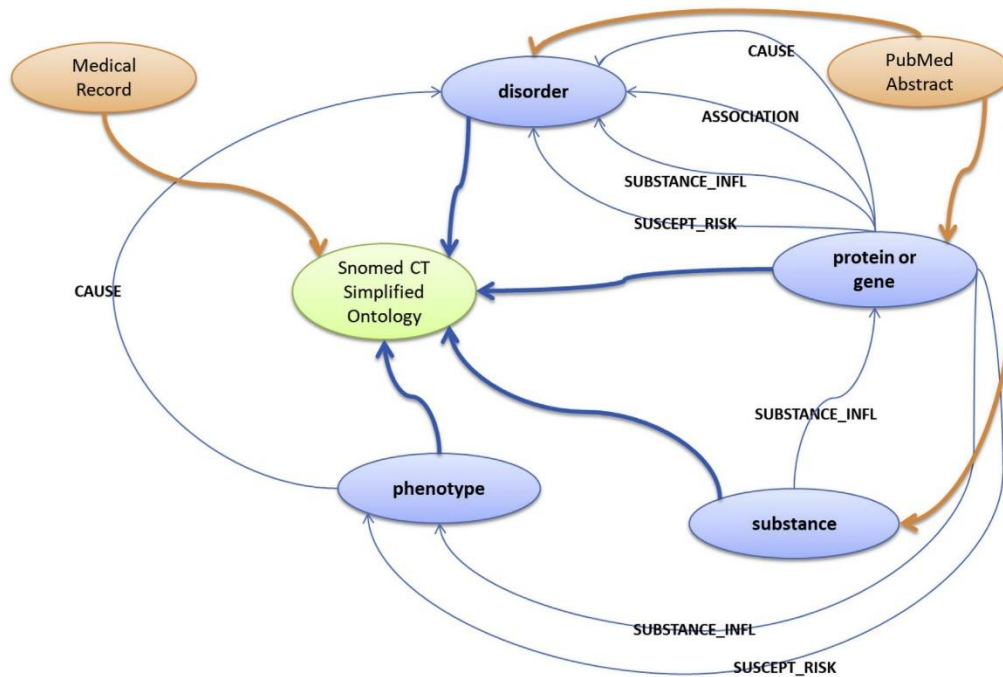


Fig. 5. Structure of the mini ontology: A graph with nodes representing entities and edges representing relations between them.

identified when its recognized name was exactly or nearly exactly as expected. We obtained a precision of 94% when counting entity repetitions (more than one mention), and 87% leaving repetitions apart.

4.3. Relation extraction

For relation extraction, we perform a manual evaluation of three relatively rare disorders: glaucoma, Wilms tumor, and pancreatitis. We need to accommodate the human effort in a timely fashion, so we focus on the obtained relations (88 in total). With such a low number, to have a first approximation to precision and recall, we consider whether the relation exists or not.

These results correspond to manual evaluation from a medical geneticist, also considering OMIM MorbidMap, which pairs each disorder with a collection of related gene symbols [43]. These first results are shown in Table 4.

Interestingly, a few novel valid relations are found, which were not registered in our OMIM MorbidMap [41], as described in Table 5.

4.3.1. Improving precision

At this point, the Relation Extraction strategy returned rather low precision figures (Table 4). We observe that many automatically inferred relations were not valid. To address this, we refine primary results by adding a Distant Supervision strategy. In our case, this is used to sharpen precision by re-classifying extracted relations as valid or invalid.

Briefly, Distant Supervision is a semi-supervised method that captures the patterns of true, known relations (usually from a known database), to segregate them from nonexistent relations. To implement Distant Supervision, we create a set of 4000 positive and 4000 negative relation samples in plain text sentences and define a set of selected features for each one. Then, we train and evaluate different classifiers,

with an evaluation set of 800 relation instances. We obtain the best results with KNN (79% precision) and SVM (74% precision).

Finally, we use the KNN classifier to filter out invalid relations from our initial results, obtaining a better precision figure at the expense of sacrificing recall (Table 6).

4.4. Graph queries

Queries to the Mini Ontology are built using Neo4J Cypher language. In the example depicted in Fig. 7, we show how a medical record is mapped to the knowledge base. In this case, the medical record includes an assertion about *hipertensión arterial* in Spanish (hypertensive disorder, or HTN, in English), which we first recognize as a known *disorder type* entity using SNOMED CT. Then, the Cypher query retrieves the subgraph of interest that applies to this disorder, composed of entities and recognized relations in PubMed.

To improve query precision, the query validates that all retrieved relations occur more than T times, where T is a safety threshold for checking that the relation exists in at least T abstract mentions. If we expand one of these relations, we can check for the abstract sentence that motivated the inference (Fig. 8). Besides this, the query ranks results by the number of evidences – a count on relations between each pair of concepts, ranking the most frequently occurring relationships in first place.

5. Discussion and conclusions

In summary, we present a framework for building software applications to help the medical geneticists diagnostics process, and accept Spanish medical records as input. We identified multiple advantages and limitations associated with our methods.

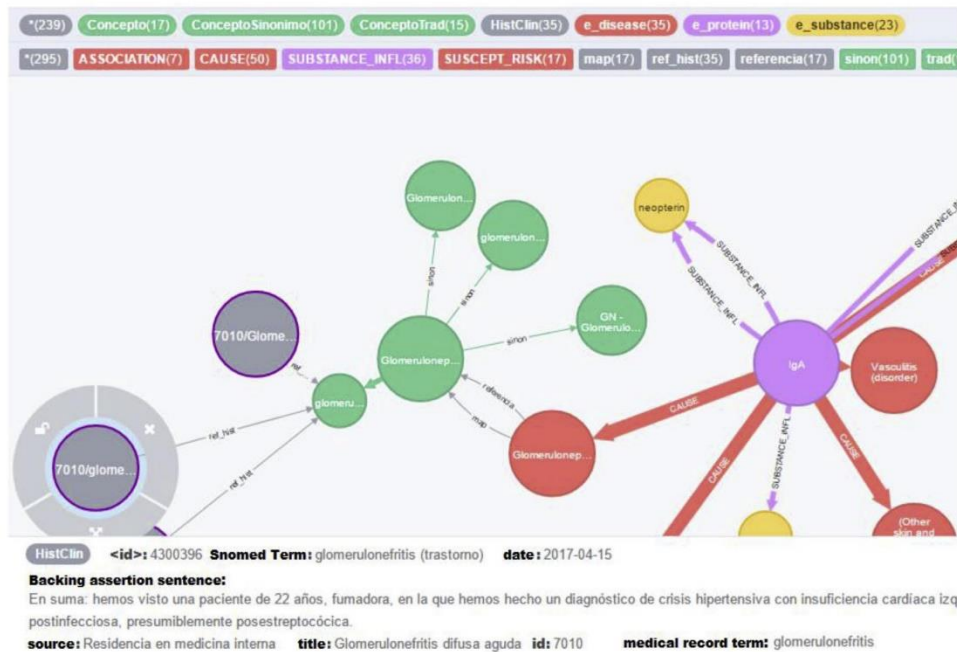


Fig. 6. Subgraph showing related entities, including medical record concepts (grey), SNOMED CT concepts [displayed as green nodes, with size big for principal concepts (FSN), medium for synonyms, and small for translations], and entities recognized in PubMed (red = disorder, yellow = substance, violet = protein/gene, pink = phenotype). ASSOCIATION, CAUSE, SUBSTANCE_INFL, and SUSCEPT_RISK are inferred relations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Precision for named entities that participate in relations in Asthma dataset.

NER Evaluation Summary – Asthma	
Retrieved Entities	4,147
True Positive	2,498
False Positive	1,649
Precision:	60%

Table 3
Precision for named entities in Spanish (medical records).

Correct or nearly correct entities, counting repetitions:	
Identified named entities:	812
Relevant entities:	760
Precision:	94%
Strict criteria (correct or nearly correct entities, not counting repetitions):	
Identified named entities:	403
Relevant entities:	351
Precision:	87%

5.1. Advantages

- **Extensibility.** The proposed workflow includes a corpus and resources set that can be extended by including more papers and ontology sources. The document acquisition part of the pipeline can integrate new sources available in multiple common ways, such as RSS feeds or REST APIs. No change is needed to the processing components that build the Mini Ontology.

Table 4
First results of relation extraction on selected rare diseases.

		Actual	
		Related	Not Related
Predicted	Related	47	41
	Not Related	0	0
		Precision =	53%
		Recall =	100%

Table 5
Automatically inferred relations, validated by manual evaluation, and not present in the OMIM MorbidMap.

Disorder	Gene Symbol	Relation
Glaucoma	Gene FOXC1	Susceptibility/Risk, Association, Cause
Wilms	Gene PCDHA	Cause
Pancreatitis	Gene HLA	Association

Table 6
Improved precision after distant supervision.

		Actual	
		Related	Not Related
Predicted	Related	20	10
	Not Related	27	31
		Precision =	67%
		Recall =	43%

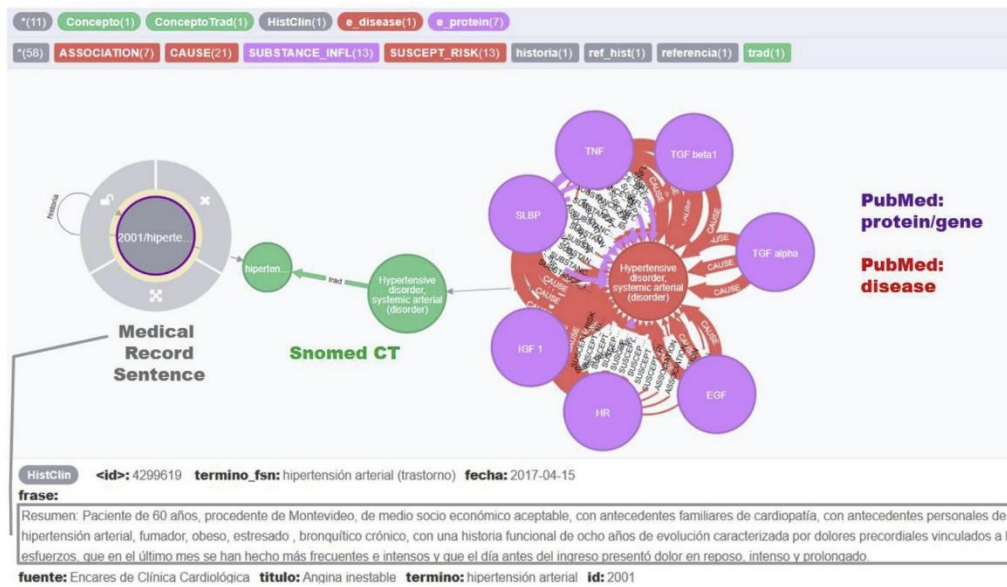


Fig. 7. Query that links a medical record to inferred relations. Each relation instance corresponds to a different mention.

- **Novel relations inference.** The framework can infer novel relations. In the evaluation of *asthma* results, we show new relations that are not present in our reference baseline (OMIM MorbidMap) but are manually evaluated as correct. This is an interesting finding since it actually corresponds to new valid relations not found in the MorbidMap release we used.
- **Precision.** Although an evaluation on a broader set is crucial, we showed how Distant Supervision is a promising method for improving precision. Even though this affects recall, the missing relations due to diminished recall can eventually be mitigated by ingesting new abstracts (or the entire document), because usually a valid relation can be found in several sentences.
- **Language independence.** The implemented pipeline addresses concepts and relation extractions from English language literature sources, and map them to Spanish medical records via SNOMED CT, which contains language agnostic concept definitions, as well as their corresponding translations. As a result, we also built a knowledge database that can be queried from external applications.
- **Advanced search features,** inherent to graphs. The Mini Ontology is implemented as a graph, which enables queries that are more complex than simple searches. Thus, instead of performing a traditional index-based search, a graph inquire can leverage the

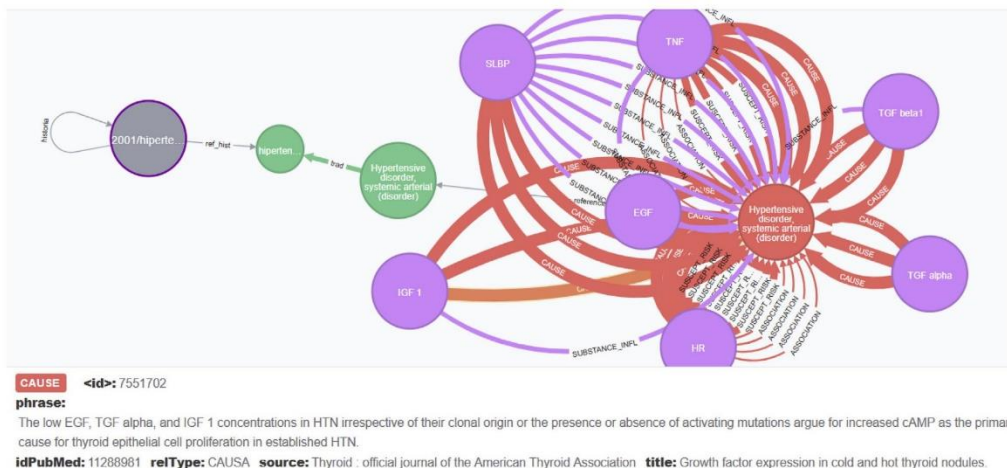


Fig. 8. Relation attributes for the selected relation (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

following capabilities:

- o **The distance between entities**, in terms of the number of relationships that connect them
- o **The centrality of concepts**, in terms of importance: how important is one concept as it relates to others
- o **Relationship strength**: how many evidences in abstracts support the existence of the inferred relationship

5.2. Limitations

- **False negatives.** We experimented with Distant Supervision for improving precision, but this could also produce a number of false negatives. Active learning could provide relief on this matter: relations that were discarded by our pipeline could be optionally marked as 'classified negative', but prompted user confirmation. This way, each originally inferred relation could be suggested for discard by our relation classifier, but the final decision could be left to the physician, thereby reducing the impact on recall.
- **Precision risks.** In our relation detection logic, good recall is prone to accepting false positives. We are aware of this situation, and as a first approach, we mitigate this by checking whether the same relation exists at different sources. But this is just a heuristic, which deserves a quantitative validation (not included in this article).

5.3. Usage notes

- At the core of our framework, the mini ontology gathers knowledge to be leveraged by consumer applications, powered by **two standard API libraries**:
 - o **Neo4J graph API**, for exploring the knowledge graph relations, using a full-fledged graph query language (Cypher). Furthermore, the graph database includes its own graph processing algorithms, such as identifying node neighborhoods (e.g. concept affinity, in our case).
 - o **Solr information retrieval API**, for querying abstracts using complex search operators. This is not a regular plain text search: since entities were previously identified (e.g. proteins, diseases), the search can focus on specific term types, or term proximity using Solr's Lucene syntax.

An application can make use of both APIs, in many different programming languages, including Java, Python and REST invocations.

5.4. Usage scenarios

- An external medical system can interrogate the graph in order to:
 - o retrieve relevant relations between causes and diseases
 - o find the articles that best suit a patient case description
 - o build custom queries for new semantic requirements, for instance:
 - determine concept affinity (e.g. neighbor proteins, in terms of semantics) – leveraging graph community detection algorithms
 - identify relationships backed by most assertions – by querying graph relation edges
- It is also possible to fulfill information retrieval requirements, which include:
 - o query for single terms in the enriched collection of abstracts – search by abstract terms, source or identified entities
 - o build complex queries on the collection, combining operators and available fields
- As a usage example that uses the above mentioned standard APIs, a chatbot prototype application was built (link to video) for a few Genetics implied disorders. This external app makes use of the Mini Ontology to answer questions like 'what is disorder X?', or 'which X factors cause Y disorder?'. The chatbot can interact in English and accept a medical record in Spanish.

Supplementary video related to this article can be found at <https://doi.org/10.1016/j.imu.2019.100181>

6. Future work

This work provides a basis for deepening into further functionalities. Below we present a list of potential future improvements:

- There are several important aspects of NLP performance that deserve more consideration. Since the framework was the main focus, it is worth noting that there is room for improvement in terms of NLP tasks, namely:
 - o Add handling of negation and hedging. An interesting work in this matter is presented in Ref. [44], and for BioNLP, in Ref. [45].
 - o Identify quantities (e.g., patient analysis measures) or time references.
 - o A number of methods could be incorporated to improve precision:
 - Dependency Parsing, for more precise relation extraction rules, combined with Semantic Role Labeling.
 - Document/Word vector representations [46], for handling document similarity and for cue words generalization.
- Broaden domain scope
 - o Incorporate new corpus sources. In particular, for Spanish and Portuguese spoken countries, we consider Scielo [47] as the main reference in Latin America.
 - o Integrate new resources, in particular, NCBI's ClinVar [48].
 - o Automate release updates on resources like thesauri and ontologies (SNOMED, OMIM, VarNomen).
 - o Extend to other biomedical subdomains, beyond Genetics and Bioinformatics.
- Extend the framework base with machine learning and neural networks platforms, such as TensorFlow [49], Theano [50], Keras [51], and Spark MLlib [52].
- Add quality metrics to inferred relations,
 - o derived from sources:
 - the impact factor of the publication
 - release date (*more recent means better*)
 - number of references to the article
 - o calculated from the graph database itself:
 - Incorporate *influence* metrics of each article (abstract), derived from graph metrics, such as *centrality*.
 - Elaborate a synthesized, summary graph, eliminating redundant relationship edges.

Funding statement

This work was supported by ICT4V - Information and Communications Technologies for Verticals (ict4v.org), grant number POS_ICT4V_2016_1_05.

Competing interests statement

We, the authors, declare that have no competing interests.

Contributorship statement

Dr. Hugo Naya conceived the presented idea. Dr. Víctor Raggio offered the medical perspective for the objectives and participated in actual results evaluations. Fernando López Bello developed the theory and performed the computations. Dr. Aiala Rosá verified the analytical methods, and encouraged Fernando López Bello to investigate improvements through applying distant supervision. All authors discussed the results and contributed to the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.imu.2019.100181>.

References

- [1] Broad Institute. Can DNA sequencing get any faster and cheaper? Broadminded Blog 2011. <https://doi.org/10.1360/zd-2013-43-6-1064>.
- [2] US National library of medicine national institutes of health, [PubMed, (n.d.)].
- [3] US National library of medicine national institutes of health, [PubMed Fact Sheet, (n.d.)].
- [4] Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Kritihara A, Petridis S, Polychronopoulos D, others. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinf* 2015;16:138.
- [5] World Health Organization (WHO). International statistical classification of diseases and related health problems. 2016.
- [6] International Health Terminology Standards Development Organisation (IHTSDO). SNOMED CT. 2017.
- [7] AGESIC. Snomed CT Uruguay. 2015.
- [8] Human Genome Variation Society (HGVS). Human variome project (HVP), HUMAN genome organization (HUGO), sequence variant nomenclature. 2017.
- [9] UniProt consortium, UniProt (universal protein resource) knowledge base, (n.d.).
- [10] UniProt, UniProtKB - Q12768 (WASC5_HUMAN), [n.d.].
- [11] Liu Y, Ge T, Mathews KS, Ji H, McGuinness DL. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. *Proc. 2015 work. Biomed. Nat. Lang. Process.* 2015. p. 92–7.
- [12] Hamon T, Graña M, Raggio V, Grabar N, Naya H. Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature. *Medinfo 2010:964–8*.
- [13] Tsuruoka Y. GENIA tagger: Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text. 2006.
- [14] Padró L, Stanilovsky E. FreeLing 3.0: towards wider multilinguality. *Proc. Lang. Resour. Eval. Conf. (LREC 2012)*. Turkey: Istanbul; 2012.
- [15] Oronoz M, Casillas A, Gojenola K, Perez A. Automatic annotation of medical records in Spanish with disease, drug and substance names. *Iberoam. Congr. Pattern recognit.* 2013. p. 536–43.
- [16] Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 2005;21:3191–2.
- [17] Hina S, Atwell E, Johnson O. Secure information extraction from clinical documents using SNOMED CT gazetteer and natural language processing. *Internet technol. Secur. Trans. (ICITST)*, 2010 Int. Conf.; 2010. p. 1–5.
- [18] Fette G, Ertl M, Wörner A, Kluegl P, Störk S, Puppe F. Information extraction from unstructured electronic health records and integration into a data warehouse. *GI-Jahrestagung*; 2012. p. 1237–51.
- [19] Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. *J Biomed Inform* 2015;58:S171–82.
- [20] Chikka VR, Mariyasagayam N, Niwa Y, Karlapalem K. Information extraction from clinical documents: towards disease/disorder template filling. *Int. Conf. Cross-language eval. Forum Eur. Lang.*; 2015. p. 389–401.
- [21] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;5:507–13.
- [22] Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A study of neural word embeddings for named entity recognition in clinical text. *AMIA annu. Symp. Proc.* 2015. p. 1326.
- [23] Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical NLP. *ACL 2016*; 2016. p. 166.
- [24] Bird S. NLTK: the natural language toolkit. *Proc. COLING/ACL Interact. Present. Sess.* 2006. p. 69–72.
- [25] Gong L. Application of biomedical text mining. *Artif. Intell. Trends appl., IntechOpen*. 2018.
- [26] Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr. ER, Mitchell TM. Toward an architecture for never-ending language learning. *AAAI*. 2010. p. 3.
- [27] Ferrucci DA. Introduction to “this is Watson. *IBM J Res Dev* 2012;56:1.
- [28] Bravo A, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinf* 2015;16:55.
- [29] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. *Proc. Jt. Conf. 47th annu. Meet. ACL 4th int. Jt. Conf. Nat. Lang. Process. AFNLP vol. 2 vol. 2*. 2009. p. 1003–11.
- [30] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypemymic propositions in biomedical text. *J Biomed Inform* 2003;36:462–77.
- [31] Kilicoglu H, Shin D, Fiszman M, Roseblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 2012;28:3158–60.
- [32] Harkema H, Roberts I, Gaizauskas R, Hepple M. Information extraction from clinical records. *Proc. 4th UK e-Science All Hands Meet.* 2005.
- [33] Campbell WS, Pedersen J, McClay JC, Rao P, Bastola D, Campbell JR. An alternative database approach for management of SNOMED CT and improved patient data queries. *J Biomed Inform* 2015;57:350–7.
- [34] Neo4j. Neo4j life sciences and healthcare network. 2017.
- [35] Berthold MR, Cebron N, Dill F, Gabriel TR, Köster T, Meinel T, Ohl P, Sieb C, Thiel K, Wiswedel B. (KNIME): the (K)onstanz (I)nfomation (M)iner. *Stud. Classif. Data anal. Knowl. Organ. (GfKL 2007)*. Springer; 2007.
- [36] Neo4j. Neo4j. 2017.
- [37] Solr. Apache Solr enterprise search. 2017.
- [38] NCBI, Entrez, [n.d.].
- [39] Apache, OpenNLP, [n.d.].
- [40] Apache, Tika, [n.d.].
- [41] McKusick VA. McKusick-nathans institute for genetic medicine. Johns Hopkins University. National Center for Biotechnology Information. National Library of Medicine. Online Mendelian Inheritance in Man, OMIM; 2004.
- [42] Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. *Proc. 52nd annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr.*; 2014. p. 55–60.
- [43] B. McKusick VA, McKusick-nathans institute for genetic medicine, Johns Hopkins University. National Center for Biotechnology Information. National Library of Medicine. Online Mendelian Inheritance in Man, OMIM: MorbidMap, [n.d.].
- [44] Monceccchi G, Minel J-L, Wonsever D. Improving speculative language detection using linguistic knowledge. *Proc. Work. Extra-propositional Asp. mean. Comput. Linguist.*. 2012. p. 37–46.
- [45] Morante R, Daelemans W. Learning the scope of hedge cues in biomedical texts. *Proc. Work. Curr. Trends biomed. Nat. Lang. Process.*. 2009. p. 28–36.
- [46] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. *Adv. Neural inf. Process. Syst.* 2013. p. 3111–9.
- [47] São Paulo Research Foundation (FAPESP), Latin American and caribbean center on health sciences information (BIREME), Scientific Electronic Library Online, SciELO, [n.d.].
- [48] NCBI, ClinVar, [n.d.].
- [49] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, others. Tensorflow: large-scale machine learning on heterogeneous distributed systems. 2016. *ArXiv Prepr. ArXiv1603.04467*.
- [50] Université de Montréal, Theano, (n.d.).
- [51] A. Keras, Keras, (n.d.).
- [52] Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai DB, Amde M, Owen S, Others, mllib: machine learning in Apache Spark. *J Mach Learn Res* 2016;17:1–7.

4. Aplicación de Supervisión Distante

4.1 Contexto

Como indicamos en el artículo publicado, a efectos de evaluar el proceso de extracción de relaciones originalmente propuesto, nos enfocamos en un conjunto de enfermedades relativamente raras: *glaucoma*, *tumor de Wilms* y *pancreatitis*. Considerando esos primeros resultados, es conveniente idear una estrategia para mejorar la precisión de la extracción. Recurrimos entonces al método *Supervisión Distante*.

Según establecen Mintz et al. [20]: “La intuición de la Supervisión Distante, es que cualquier frase que contenga un par de entidades que participan en una relación conocida [...] es probable que exprese esa relación de alguna manera”. Es una estrategia de aprendizaje de relaciones que construye un modelo de aprendizaje automático a partir de una verdad conocida, con relaciones previamente anotadas, para clasificar una relación candidata como existente o inexistente.

Este método requiere evaluar múltiples características extraídas de la oración en lenguaje natural, para lo cual se construyen modelos de aprendizaje automático, que intentan inferir el patrón que ellas contienen. DeepDive [25] fue un esfuerzo por automatizar esta tarea, proponiendo una plataforma genérica de aprendizaje basado en Supervisión Distante a partir de ejemplos suministrados por un usuario. Se han obtenido buenos resultados con modelos más específicos, como los que propone Greaves [26], basados en SVM.

Para el dominio BioNLP Liu et al. [27] desarrollan un método de supervisión distante, aprovechando el recurso SemRep [28]. Tomando como línea base de comparación enfoques previos de SVMs transductivas (variante de SVM que puede utilizar datos parcialmente etiquetados), mejoran la performance de clasificación de relaciones extraídas, e inclusive extienden la definición de probabilidad de relación a nivel corpus, más allá de la oración.

4.2 Estrategia

Las relaciones que obtiene nuestro proceso de extracción original son del estilo:

```
<tipo de relación, entidad nombrada 1, entidad nombrada 2>
```

La idea es **filtrar** las relaciones incorrectas, y así lograr una mejor precisión.

Utilizando la base de datos OMIM, que correlaciona símbolos de genes/proteínas y trastornos en una relación causal/asociativa genérica, entrenamos diferentes clasificadores para detectar si existe o no una relación candidata. Nuestra referencia para la evaluación de la corrección fue OMIM MorbidMap, que combina cada trastorno con una colección de símbolos genéticos relacionados. Esta es nuestra definición de *ground truth* para el problema de la clasificación, donde tendremos relaciones válidas estilo:

<gen, fenotipo>

Vale notar que OMIM MorbidMap no necesariamente tipifica la relación: sólo informa que existe algún tipo de relación entre un gen/proteína y un fenotipo. Por esta razón, no es posible mejorar la clasificación de relaciones obtenida desde las reglas.

4.3 Implementación

En nuestro caso, la Supervisión Distante es una función:

$$\text{dist_sup}(K, v) \rightarrow \text{Boolean}$$
$$v = \{a_1, a_2, \dots, a_n\}$$

donde K es el clasificador, y v es el vector de atributos observados en la oración. Estos atributos se generan para cada una de las oraciones procesadas, e incluyen los siguientes conceptos:

- **Tipo de Relación** inferida de la oración, a partir del enfoque basado en reglas que comentamos anteriormente. En el siguiente ejemplo, se infiere el tipo de relación CAUSA:

The aim of our study is to investigate a potential involvement of LOXL1 gene in the pathogenesis of pigment dispersion syndrome (PDS) and pigmentary glaucoma (PG).

Relación encontrada: **CAUSA** definida por: gene-symbol = loxl1 (GENE~SYMBOL~LOXL1) / disease=pigment dispers syndrom (DISEASE~pigment dispersion syndrome) / cue=involv

- **Palabra señal (cue)** que se indicó en las reglas de anotación automática. En el ejemplo anterior, involvement.
- **Lema (lemma)** de una palabra es su forma canónica, esto es, la que habitualmente se encuentra en un diccionario. Ejemplo: palabra diagnosing, lema diagnose.
- **Raíz (stem)** de una palabra. En morfología lingüística, la raíz es la forma base de una palabra, obtenida por un algoritmo de *stemming*. Las diferentes formas morfológicas se derivan a partir de esta raíz. Por ejemplo, la raíz de genetic es genet.
- El **etiquetado gramatical (Part Of Speech tagging)** asigna categorías gramaticales (*POS tags*) a las palabras de un texto. Para el ejemplo anteriormente mencionado, el etiquetado gramatical es:

The aim of our study is to investigate a potential involvement of LOXL1 gene in the pathogenesis of pigment dispersion syndrome (PDS) and pigmentary glaucoma (PG) .

Adjective
Adverb
Conjunction
Determiner
Noun
Number
Preposition
Pronoun
Verb

Más específicamente, tomamos los siguientes atributos:

- Tipo de relación
- Tipo de Entidad Nombrada 1
- Tipo de Entidad Nombrada 2
- Raíz de la palabra Señal
- Posición relativa de Señal vs. Entidades Nombradas
- Distancia (cantidad de palabras) entre Entidades Nombradas
- Para EN1: POS tag de palabra-2
- Para EN1: lema de la palabra-2
- Para EN1: POS tag de palabra-1
- Para EN1: lema de la palabra-1
- Para EN1: POS tag de palabra+1
- Para EN1: lema de la palabra+1
- Para EN1: POS de palabra+2
- Para EN1: lema de la palabra+2
- Para EN2: POS tag de palabra-2
- Para EN2: lema de la palabra-2
- Para EN2: POS tag de palabra-1
- Para EN2: lema de la palabra-1
- Para EN2: POS tag de palabra+1
- Para EN2: lema de la palabra+1
- Para EN2: POS tag de word+2
- Para EN2: lema de la palabra+2

Ejemplo:

Oración	Patients with the Denys Drash syndrome (Wilms tumor , genital anomalies, and nephropathy) have been demonstrated to carry de novo constitutional mutations in WT1 , the Wilms tumor gene at chromosome 11p13.	
<i>Valor de la Entidad Nombrada 1</i>	ValNE1	WT1
<i>Valor de la Entidad Nombrada 2</i>	ValNE2	Wilms tumor
Atributos:		
<i>Tipo de Relación</i>	TipoRel	CAUSA
<i>Tipo de Entidad Nombrada 1</i>	TNE1	GENE-SYMBOL
<i>Tipo de Entidad Nombrada 2</i>	TNE2	DISEASE
<i>Raíz de la palabra Señal</i>	cueStem	mutat
<i>Posición de Señal vs. Entidades Nombradas</i>	poscue	2
<i>Distancia entre Entidades Nombradas</i>	distent	15
<i>Para EN1: lema de palabra-2</i>	1_ant2lemma	mutation
<i>Para EN1: POS de la palabra-2</i>	1_ant2pos	NNS
<i>Para EN1: lema de palabra-1</i>	1_ant1lemma	in
<i>Para EN1: POS de la palabra-1</i>	1_ant1pos	IN
<i>Para EN1: lema de palabra+1</i>	1_dsp1lemma	,
<i>Para EN1: POS de la palabra+1</i>	1_dsp1pos	Fc
<i>Para EN1: lema de palabra+2</i>	1_dsp2lemma	the
<i>Para EN1: POS de la palabra+2</i>	1_dsp2pos	DT
<i>Para EN2: lema de word-2</i>	2_ant2lemma	syndrome
<i>Para EN2: POS de la palabra-2</i>	2_ant2pos	NN
<i>Para EN2: lema de palabra-1</i>	2_ant1lemma	(
<i>Para EN2: POS de la palabra-1</i>	2_ant1pos	Fpa
<i>Para EN2: lema de palabra+1</i>	2_dsp1lemma	,
<i>Para EN2: POS de la palabra+1</i>	2_dsp1pos	Fc
<i>Para EN2: lema de word+2</i>	2_dsp2lemma	genital
<i>Para EN2: POS de la palabra+2</i>	2_dsp2pos	JJ

Es importante notar que una sola oración podría tener N pares de entidades, lo que llevaría a N diferentes vectores de atributos ∇ .

4.4 Preparación de Datos para Entrenamiento

Para nuestro conjunto de datos de entrenamiento, consideramos oraciones de abstracts de PubMed, donde ya realizamos previamente la extracción de relaciones, donde el resumen menciona al menos una de las enfermedades en consideración. A partir de ahí, definimos un *subconjunto de frases positivas* (la relación existe, de acuerdo a nuestro ground truth) y otro de *negativas* (la relación no existe).

- **Subconjunto de oraciones positivas**

En estas frases hay evidencia de una relación válida, según un par OMIM MorbidMap $\langle \text{gen}, \text{fenotipo} \rangle$. Tomamos una muestra de 4000 ocurrencias.

- **Subconjunto de oraciones negativas**

Uno de los retos que plantea Supervisión Distante, es encontrar muestras negativas adecuadas [20], [26]. Es frecuente encontrar que el conjunto de muestra sea muy desbalanceado hacia los negativos, con lo cual una simple selección al azar puede ser una buena primera aproximación.

En nuestro caso, siguiendo nuestros criterios definidos, interpretamos "negativa" a una relación que no puede ser verificada contra las existentes en OMIM MorbidMap. Nuevamente, tomamos una muestra de 4000 ocurrencias.

Estas 8000 muestras serán la semilla para construir y probar diferentes clasificadores \mathcal{K} (90% para entrenamiento, 10% para evaluación).

Métodos de Aprendizaje Automático

Se consideran diferentes familias de clasificadores [29] para nuestro problema:

Árbol de clasificación. Uno de los primeros métodos de clasificación, que ofrece un modelo simple y una fácil interpretación. Ejemplos: Árboles de Clasificación y Regresión (CART), C4.5.

Vecinos más cercanos. Uno de los ejemplos más conocidos es K vecinos más cercanos (K-NN). En esta familia de métodos, se utiliza una medida de disimilitud (δ). Para la construcción de δ , la optimización matemática es relevante. δ puede ser una función simple (por ejemplo, la distancia euclidiana) o más compleja (por ejemplo, la distancia de edición). Otras métricas pueden ser consideradas, como la distancia de Mahalanobis. La elección de δ condiciona el rendimiento de K-NN.

Clasificador lineal. Los datos se asumen en un espacio vectorial y se intenta construir el mejor separador lineal posible entre clases. Ejemplos: Clasificador lineal de Fisher, análisis discriminante lineal (LDA), regresión logística y heurística basada en redes neuronales como el perceptrón. En particular, el perceptrón con múltiples capas ocultas (**MLP**) es un modelo de aprendizaje profundo.

Support vector machine. Estos métodos tratan de encontrar el mejor hiperplano para separar (clasificar) los puntos n-dimensionales en clases inconexas. El SVM tiene varias virtudes: se basa en una teoría intuitiva, requiere pocos ejemplos para su entrenamiento y no es sensible al número de dimensiones. Una función de núcleo se utiliza para mapear el espacio de muestreo a un espacio

de mayor dimensión, para un mejor aprovechamiento de la potencia computacional. Una de las funciones del núcleo más utilizadas es la función de base radial (RBF).

De esos grupos, entrenamos y comparamos las siguientes opciones:

- Árbol de decisión
- K vecinos más cercanos (KNN)
- Regresión logística
- Red neuronal de tipo perceptrón multicapa (MLP)
- Support vector machine (SVM)

4.5 Resultados

Presentamos a continuación algunas métricas de performance al aplicar Supervisión Distante al filtro de relaciones válidas, ensayando diferentes métodos de aprendizaje automático. En primer término, veremos la performance del método per se, y finalmente, cómo mejora la performance respecto al método de evaluación ya realizado.

Consideramos las siguientes métricas de evaluación:

- **Precisión:** ratio de relaciones identificadas que son correctas (hay relación), sobre el total de relaciones identificadas (sean válidas o no).
$$\text{Precisión} = TP / (TP + FP)$$
- **Cobertura (*recall*):** ratio de relaciones identificadas que son correctas (hay relación), sobre el total de relaciones válidas (hayan sido identificadas o no).
$$\text{Cobertura} = TP / (TP + FN)$$
- **Medida F1:** balance entre precisión y cobertura, calculado como media armónica entre ambos.
$$F1 = 2 * \text{Precisión} * \text{Cobertura} / (\text{Precisión} + \text{Cobertura})$$

4.5.1 Entrenamiento para Diferentes Métodos de Aprendizaje Automático

La Tabla 2 enumera los resultados para cada método de aprendizaje automático que ensayamos, sobre una muestra de 800 relaciones candidatas:

Método	Cobertura	Precisión	Medida F1	Notas
Árbol de Decisión	52%	60%	56%	Se consideraron las predicciones indecibles por el modelo, como "No hay relación"
KNN	91%	82%	86%	Conversión de categorías a números y K = 15
Regresión Logística	89%	65%	76%	Se usó Gradiente Promedio Estocástico
Red Neuronal – Perceptrón Multicapa	69%	65%	67%	Simétrica, 2 capas ocultas, 20 neuronas por capa
SVM (RBF)	93%	76%	83%	Se usó función kernel RBF (mejor que polinómica y tangente hiperbólica)

Tabla 2 – Performance de métodos de aprendizaje automático utilizados en Supervisión Distante.

4.5.2 Incidencia en la Performance de Extracción de Relaciones

Realizamos una evaluación sobre las relaciones extraídas para los casos de las tres enfermedades raras seleccionadas previamente. Notoriamente, es un conjunto reducido de relaciones, por lo que debe considerarse como una primera aproximación, sobre la que se podría trabajar en mayor extensión, con enfermedades más comunes.

La Tabla 3 muestra la evaluación del rendimiento de la extracción de relaciones antes y después de aplicar el filtro de Supervisión Distante. Obsérvese cómo ha mejorado la precisión para las tres enfermedades seleccionadas.

Evaluation before Distant Supervision

Evaluation																								
Overall							Glaucoma						Pancreatitis						Wilms					
Actual							Actual						Actual						Actual					
ASSOC CAUSE INFSUS NO REL SURI							ASSOC CAUSE INFSUS NO REL SURI						ASSOC CAUSE INFSUS NO REL SURI						ASSOC CAUSE INFSUS NO REL SURI					
Predicted	ASSOC	10			7		ASSOC	2			5		ASSOC	6			1		ASSOC	2			1	3
	CAUSE		26		23		CAUSE		3		8		CAUSE		17		15		CAUSE		6			6
	INFSUS			2	2		INFSUS						INFSUS			2			INFSUS			2		2
	NO REL						NO REL						NO REL						NO REL					
	SURI				9	9	SURI				5	2	SURI			2	6		SURI			2		3
Measures																								
	TP	TotPred	TotAct	Prec	Rec	F1	TP	TotPred	TotAct	Prec	Rec	F1	TP	TotPred	TotAct	Prec	Rec	F1	TP	TotPred	TotAct	Prec	Rec	F1
ASSOC	10	17	10	0,59	1,00	0,74	2	7	2	0,29	1,00	0,44	6	7	6	0,86	1,00	0,92	2	6	2	0,33	1,00	0,50
CAUSE	26	49	26	0,53	1,00	0,69	3	11	3	0,27	1,00	0,43	17	32	17	0,53	1,00	0,69	6	12	6	0,50	1,00	0,67
INFSUS	2	4	2	0,50	1,00	0,67	0	0	0				0	2	0	0,00			2	4	2	0,50	1,00	0,67
NO REL	0	0	41		0,00		0	0	18		0,00		0	0	20		0,00		0	0	3		0,00	
SURI	9	18	9	0,50	1,00	0,67	2	7	2	0,29	1,00	0,44	6	8	6	0,75	1,00	0,86	3	5	14	0,60	0,21	0,32
General	47	88	47	0,53	1,00	0,70	7	25	7	0,28	1,00	0,44	29	49	29	0,59	1,00	0,74	13	27	24	0,48	0,54	0,51

Evaluation after Distant Supervision

Evaluation																								
Overall							Glaucoma						Pancreatitis						Wilms					
Actual							Actual						Actual						Actual					
ASSOC CAUSE INFSUS NO REL SURI							ASSOC CAUSE INFSUS NO REL SURI						ASSOC CAUSE INFSUS NO REL SURI						ASSOC CAUSE INFSUS NO REL SURI					
Predicted	ASSOC	5			3		ASSOC	2			3		ASSOC	2			2		ASSOC	1				
	CAUSE		9		3		CAUSE		3		1		CAUSE		4		2		CAUSE		2			
	INFSUS			1			INFSUS						INFSUS						INFSUS			1		
	NO REL	5	17	1	31	4	NO REL			11			NO REL	4	13		17	4	NO REL	1	4	1	3	
	SURI				4	5	SURI			3	2		SURI			1	2		SURI					1
Measures																								
	TP	TotPred	TotAct	Prec	Rec	F1	TP	TotPred	TotAct	Prec	Rec	F1	TP	TotPred	TotAct	Prec	Rec	F1	TP	TotPred	TotAct	Prec	Rec	F1
ASSOC	5	8	10	0,63	0,50	0,56	2	5	2	0,40	1,00	0,57	2	2	6	1,00	0,33	0,50	1	1	2	1,00	0,50	0,67
CAUSE	9	12	26	0,75	0,35	0,47	3	4	3	0,75	1,00	0,86	4	6	17	0,67	0,24	0,35	2	2	6	1,00	0,33	0,50
INFSUS	1	1	2	1,00	0,50	0,67													1	1	2	1,00	0,50	0,67
NO REL	31	58	41	0,53	0,76	0,63	11	11	18	1,00	0,61	0,76	17	38	20	0,45	0,85	0,59	3	9	3	0,33	1,00	0,50
SURI	5	9	9	0,56	0,56	0,56	2	5	2	0,40	1,00	0,57	2	3	6	0,67	0,33	0,44	1	1	1	1,00	1,00	1,00
General	20	30	47	0,67	0,43	0,52	7	14	7	0,50	1,00	0,67	8	11	29	0,73	0,28	0,40	5	5	11	1,00	0,45	0,63

Tabla 3 - - Evaluación del rendimiento de la extracción de relaciones antes y después del refinamiento aplicando el método de Supervisión Distante. TP=Positivos verdaderos, TotPred=Total previsto, TotAct=Total real, Prec=Precisión, Rec=Cobertura (Recall), F1=Medida F1

En primer lugar, antes de aplicar Supervisión Distante, obtenemos una precisión General (agregada de los tres trastornos) de 0.53, especialmente pobre para el caso de Glaucoma (0.28). La cobertura obtenida es alta; casi todas las relaciones potenciales son capturadas. Al tener bajas cifras de precisión, el problema aquí está relacionado con la falta de exactitud o, peor aún, con la inexistencia biológica de relación(es) inferida(s). Desde un perspectiva médica esto es especialmente problemático.

La segunda parte de la evaluación se realiza después de aplicar el método de Supervisión Distante, filtrando las relaciones descartadas por el clasificador. El aumento de la precisión (0.67 para el caso General, 0.50 para Glaucoma, 0.73 para Pancreatitis y 1.00 para el Wilms) se produce a costa de reducir

la cobertura (0.43 para el caso General). Dependiendo del contexto funcional, esta precisión superior podría ser considerada como una mejora en términos de significación (por ejemplo en caso de un diagnóstico), al asegurar mejor la identificación de relaciones verdaderas.

5. Consideraciones Finales

Hemos construido un marco de trabajo para acercar los hallazgos académicos en medicina genética, a textos provenientes de historias clínicas, integrando diversos recursos léxicos en inglés y español. Nuestra propuesta incluye una ontología que aprovecha los resultados de esta integración, agregando relaciones nuevas que se infieren partir de las fuentes.

Por otra parte, logramos un avance en la performance de nuestro proceso de extracción de relaciones, al aplicar la estrategia basada en Supervisión Distante. Más específicamente, la métrica de precisión es de particular importancia en el dominio médico.

Como se indica en las secciones 5 y 6 del artículo, hemos identificado algunas mejoras y posibles cursos de acción a futuro sobre la base de lo implementado. En este apartado, tomaremos nota de algunas referencias adicionales que analizamos con posterioridad a la publicación.

Desde el inicio de este trabajo hasta su publicación, han surgido propuestas de abordaje a la extracción de información desde historias clínicas que aprovechan recursos que representan el estado del arte en aprendizaje máquina, con mediciones de performance que desplazan el uso de aproximaciones más clásicas basadas en reglas.

En diversas revisiones de la literatura disponible y métodos aplicados a la fecha [30], [31], se enumeran algunos elementos que hemos tenido en cuenta en nuestra propuesta:

- métodos de clasificación como SVM,
- de reconocimiento de entidades como CRF, y
- SNOMED-CT como fuente terminológica de consenso

En una línea similar a la de este trabajo, CLAMP [32] propone un toolkit para construcción de pipelines en textos clínicos, que agrega una interfaz de usuario para realizar anotaciones. Si bien se trata de una iniciativa más abarcativa, que contempla la participación del usuario final, es una referencia de interés para comparar decisiones de diseño de nuestro trabajo.

Por otro lado, hay buenas referencias recientes de performance de uso de redes neuronales de tipo Long Short Term Memory (LSTM) para reconocimiento de menciones de genes [33]. El uso de este tipo de red neuronal viene en aumento en tareas de PLN en el último lustro, y es una alternativa a considerar para reconocimiento de entidades. En la misma línea, otro trabajo [34] utiliza LSTMs para extracción de relaciones, restringidas –como en nuestro trabajo- al ámbito de una oración. El pipeline que hemos presentado en el artículo, admitiría la inclusión de LSTMs, por ejemplo utilizando Keras y Tensorflow en KNIME.

Una tarea pendiente en el marco de trabajo propuesto, es el tratamiento de frases especulativas y negaciones. Para el caso de negación en historias clínicas en español, Santiso et al. [35] presentan una adaptación de la herramienta NegEx, con resultados alentadores.

En cuanto a búsqueda de similaridad entre términos biomédicos, Zhu et al. [36] exploran el efecto de variables como antigüedad y tamaño, al utilizar *embeddings*, en este caso Word2vec.

Además del trabajo futuro expresado en la sección 6 del artículo, vale la pena considerar la posibilidad de extender este pipeline de procesamiento a nuevos textos, incluso adicionales a la fuente PubMed, procesando la totalidad del cuerpo del artículo y no solamente los abstracts. Los artículos de texto completo en general incluyen enunciados más largos y complejos que los abstracts, tornando aun más necesario al abordaje de las limitaciones expresado en esa sección (negaciones, especulación, extracción de variables cuantitativas y temporales).

Bibliografía

- [1] US National Library of Medicine National Institutes of Health, «PubMed». [En línea]. Disponible en: <https://www.ncbi.nlm.nih.gov/pubmed/>. [Accedido: 30-dic-2018].
- [2] Wikipedia, «BioNLP». [En línea]. Disponible en: https://en.wikipedia.org/wiki/Biomedical_text_mining. [Accedido: 18-abr-2019].
- [3] U. Georgios Paliouras (NCSR «Demokritos», Greece and University of Houston, U. Prof. Ioannis A. Kakadiaris (University of Houston, y G. Anastasia Krithara (NCSR «Demokritos», «BioASQ - A challenge on large-scale biomedical semantic indexing and question answering». [En línea]. Disponible en: <http://bioasq.org/project>. [Accedido: 05-jul-2018].
- [4] A. Kosmopoulos, I. Androutsopoulos, y G. Paliouras, «Biomedical semantic indexing using dense word vectors in bioasq», *J. Biomed. Semant. Suppl. Semant. Biomed. Inf. Retr.*, pp. 5-7, 2015.
- [5] NCBI, «Entrez». [En línea]. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.The_Entrez_Databases. [Accedido: 31-dic-2017].
- [6] International Health Terminology Standards Development Organisation (IHTSDO), «Snomed CT», 2017. [En línea]. Disponible en: <http://www.snomed.org>. [Accedido: 11-nov-2018].
- [7] AGESIC, «Snomed CT Uruguay», 2015. [En línea]. Disponible en: <https://www.agesic.gub.uy/innovaportal/v/4829/1/agesic/primer-release-de-la-extension-uruguay-de--snomed-ct®.html>. [Accedido: 05-abr-2017].
- [8] Wikipedia, «Snomed CT», *Wikipedia*. [En línea]. Disponible en: https://en.wikipedia.org/wiki/SNOMED_CT. [Accedido: 01-abr-2019].
- [9] International Health Terminology Standards Development Organisation (IHTSDO), «SNOMED CT Browser». [En línea]. Disponible en: <https://browser.ihtsdotools.org>. [Accedido: 07-may-2019].
- [10] Wikipedia, «Snomed CT statistics». [En línea]. Disponible en: https://en.wikipedia.org/wiki/SNOMED_CT. [Accedido: 12-dic-2017].
- [11] B. McKusick VA, McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University. National Center for Biotechnology Information, National Library of Medicine, «Online Mendelian Inheritance in Man, OMIM», 2004. .
- [12] Oficina del Libro (FMED), *Encares de Clínica Neurológica 1*. .
- [13] Oficina del Libro (FMED), *Encares de Clínica Neurológica 2*. .
- [14] Oficina del Libro (FMED), *Encares para el Internado Obligatorio*. .
- [15] Oficina del Libro (FMED), *Hematología*. .
- [16] Oficina del Libro (FMED), *Residencia en Medicina Interna*. .

- [17] Oficina del Libro (FMED), *Encares de Clínica Cardiológica*. .
- [18] B. Settles, «ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text», *Bioinformatics*, vol. 21, n.º 14, pp. 3191-3192, 2005.
- [19] S. et al., «BioCreAtIvE II GM corpus (Gene Mention task)», 2008. [En línea]. Disponible en: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2559986>. [Accedido: 11-nov-2018].
- [20] M. Mintz, S. Bills, R. Snow, y D. Jurafsky, «Distant supervision for relation extraction without labeled data», en *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 2009, pp. 1003-1011.
- [21] B. McKusick VA, McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University. National Center for Biotechnology Information, National Library of Medicine, «Online Mendelian Inheritance in Man, OMIM: MorbidMap». [En línea]. Disponible en: https://www.omim.org/help/faq#1_5. [Accedido: 25-may-2017].
- [22] M. R. Berthold *et al.*, «{KNIME}: The {K}onstanz {I}nformation {M}iner», en *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, 2007.
- [23] Neo4j, «Neo4j Graph Database», 2017. [En línea]. Disponible en: <https://neo4j.com/>. [Accedido: 11-nov-2018].
- [24] L. Padró y E. Stanilovsky, «FreeLing 3.0: Towards Wider Multilinguality», en *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, 2012.
- [25] Niu et al., «DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference», *VLDS*, vol. 12, pp. 25-28, 2012.
- [26] M. W. Greaves, «Relation Extraction using Distant Supervision, SVMs, and Probabilistic First Order Logic», Carnegie Mellon University, 2014.
- [27] R. Liu, M., Ling, Y., An, Y., Hu, X., Yagoda, A., & Misra, «Relation Extraction from Biomedical Literature with Minimal Supervision and Grouping Strategy», en *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 444-449.
- [28] T. C. Rindflesch y M. Fiszman, «The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text», *J. Biomed. Inform.*, vol. 36, n.º 6, pp. 462-477, 2003.
- [29] D. R. Carrizosa, E., & Morales, «Supervised classification and mathematical optimization», *Comput. Oper. Res.*, vol. 40, n.º 1, pp. 150-165, 2013.
- [30] S. and others Wang, Yanshan and Wang, Liwei and Rastegar-Mojarad, Majid and Moon, Sungrim and Shen, Feichen and Afzal, Naveed and Liu, Sijia and Zeng, Yuqun and Mehrabi, Saeed and Sohn, «Clinical information extraction applications: a literature review», *J. Biomed. Inform.*, vol. 77, pp. 34-49, 2018.
- [31] T. Kreimeyer, Kory and Foster, Matthew and Pandey, Abhishek and Arya, Nina and Halford, Gwendolyn and Jones, Sandra F and Forshee, Richard and Walderhaug, Mark and Botsis, «Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review», *J. Biomed. Inform.*, vol. 73, pp. 14--29, 2017.

- [32] H. Soysal, Ergin and Wang, Jingqi and Jiang, Min and Wu, Yonghui and Pakhomov, Serguei and Liu, Hongfang and Xu, «CLAMP--a toolkit for efficiently building customized clinical natural language processing pipelines», *J. Am. Med. Informatics Assoc.*, vol. 25, n.º 3, pp. 331--336, 2017.
- [33] D. Lyu, Chen and Chen, Bo and Ren, Yafeng and Ji, «Long short-term memory RNN for biomedical named entity recognition», *BMC Bioinformatics*, vol. 18, n.º 1, p. 462, 2017.
- [34] D. Li, Fei and Zhang, Meishan and Fu, Guohong and Ji, «A neural joint model for entity and relation extraction from biomedical text», *BMC Bioinformatics*, vol. 18, n.º 1, p. 198, 2017.
- [35] M. Santiso, Sara and Casillas, Arantza and Pérez, Alicia and Oronoz, «Medical entity recognition and negation extraction: assessment of NegEx on health records in Spanish», en *International Conference on Bioinformatics and Biomedical Engineering*, 2017, pp. 177-178.
- [36] F. Zhu, Y., Yan, E., & Wang, «Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec», *BMC Med. Inform. Decis. Mak.*, vol. 17, p. 95, 2017.