LitCoin NLP Challenge: Part 1

Identificación de menciones de entidades biomédicas en artículos científicos

Universidad del Valle
de Guatemala
Data Science - Sección 20
Proyecto 1
Análisis de datos

Integrantes:
Julio Herrera
Guido Padilla
Diego Álvarez
Oscar Paredez

Problemática

Gran cantidad de artículos científicos con menciones a entidades biomédicas de distintas formas usando sinónimos, abreviaturas, descripciones, etc.

SCIENTIFIC ARTICLES STRUCTURE[1]

Parlindungan Pardede
Universitas Kristen Indonesia Jakarta
e-mail: parlpard2010@gmail.com

Abstract
Scientists and researchers communicate their research results one to another through scientific articles. These articles are generally published in scientific journals or presented in conferences. To make the communication efficient and effective, the articles must be presented coherently and logically. This can be realized through the use of the format commonly used in scientific articles. This paper describes the structure of scientific articles that are commonly used to communicate the results of research, known as AIMReDCaR (Abstract, Introduction, Methodology, Result, Conclusion, and References). Discussions are focused on the scientific article features and guidelines for writing each section.

Keywords: *AIMReDCaR*, scientific article, journals

Introduction

Scientific articles are the 'storehouses' of scientific researches results plus the procedures used to make those researches. They are written to provide a means for scientists to communicate each other about the results of their researches. To make the communication effective, the media (manuscripts) must have a standardized framework so that the authors could present their findings and ideas in an orderly, logical manner. This paper introduces the generic structure of scientific articles written based on actual and relevant studies. Discussions are focused on

Conceptos del proyecto

Entidad biomédica: Términos significativos en el ámbito biológico, como genes, enfermedades, sustancias químicas y nombres de medicamentos.

Ontología: Define una entidad biomédica a partir del significado de los datos que la componen y las asociaciones entre ella.

Asociaciones: Se dan entre distintas entidades. Usa palabras conocidas para encontrarlas como:

- … abundancia afectada por …
- … actúa antes de …
- … degradación afectada por …

Entity Recognizer

NER es un modelo parte de la librería de spaCy, que spaCy es la librería para Natural Language Processing. NER contiene diferentes corpus que se usan en conjunto para la detección de palabras en el contexto de entidades biomédicas.

- Craft
- JNLPBA
- BC5CDR
- BIONLP

El corpus es la colección de textos nativos, es decir, escritos por un humano y que son usados como referencia en un contexto general o específico para el entrenamiento en detección de palabras para NLP.

Ejemplo de observación de entidad

abstract_id	title	abstract
1353340	Late-onset metachromatic leukodystrophy: molecular pathology in two siblings.	We report on
1671881	Two distinct mutations at a single BamHI site in phenylketonuria.	Classical phe
1848636	Debrisoquine phenotype and the pharmacokinetics and beta-2 receptor pharmacodynamics of metoprolol and its enantiomers.	The metaboli
2422478	Midline B3 serotonin nerves in rat medulla are involved in hypotensive effect of methyldopa.	Previous exp
2491010	Molecular and phenotypic analysis of patients with deletions within the deletion-rich region of the Duchenne muscular dystrophy (DMD) gene.	Eighty unrela
7468724	Cardiovascular complications associated with terbutaline treatment for preterm labor.	Severe cardic

	id	abstract_id	offset_start	offset_finish	type	mention	entity_ids
0	0	1353340	11	39	DiseaseOrPhenotypicFeature	metachromatic leukodystrophy	D007966
1	1	1353340	111	126	GeneOrGeneProduct	arylsulfatase A	410
2	2	1353340	128	132	GeneOrGeneProduct	ARSA	410
3	3	1353340	159	187	DiseaseOrPhenotypicFeature	metachromatic leukodystrophy	D007966
4	4	1353340	189	192	DiseaseOrPhenotypicFeature	MLD	D007966
5	5	1353340	210	220	SequenceVariant	arginine84	rs74315458

Resultados

- Luego de utilizar y analizar textos haciendo referencia a artículos científicos podemos observar las menciones de entidades en cada uno de estos, en base a la palabra, posición inicial y final.

	text	label_	start_char	end_char
0	thiazide-sensiti	GeneOrGeneProduct	29	67
1	patients	OrganismTaxon	86	93
2	Gitelman	DiseaseOrPhenotypicFeature	100	107
3	Gitelman	DiseaseOrPhenotypicFeature	133	140
4	GS	GeneOrGeneProduct	154	155
5	thiazide-sensiti	GeneOrGeneProduct	238	276
6	patients	OrganismTaxon	492	499
7	GS	GeneOrGeneProduct	506	507
8	Peripheral blooc	<NA>	519	552
9	patients	OrganismTaxon	573	580

Hallazgos y Conclusiones

- Las entidades identificadas por los distintos modelos varían, según como fueron pre entrenados y la cantidad que estos contienen.
- Se observó que en algunos casos las entidades no tienen relación con las palabras analizadas, sin embargo esto no era común.

GRACIAS !!!