El efecto de la especificidad de los datos de entrenamiento de modelos de grafo de conocimiento: un estudio de biología vegetal molecular

Serena G. Lotreck[1,2], Shin-han Shiu[1,2]

1. Department of Plant Biology, Michigan State University, East Lansing, MI 48824
2. Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824

ABSTRACT

Grafos de conocimiento es una técnica del procesamiento de lenguas naturales que puede ser muy útil en el ámbito biológico para extraer información automáticamente de artículos científicos y generar hipótesis. Este proyecto explora el efecto de la especificidad de los datos de entrenamiento de los modelos de extracción de información en la biología vegetal.

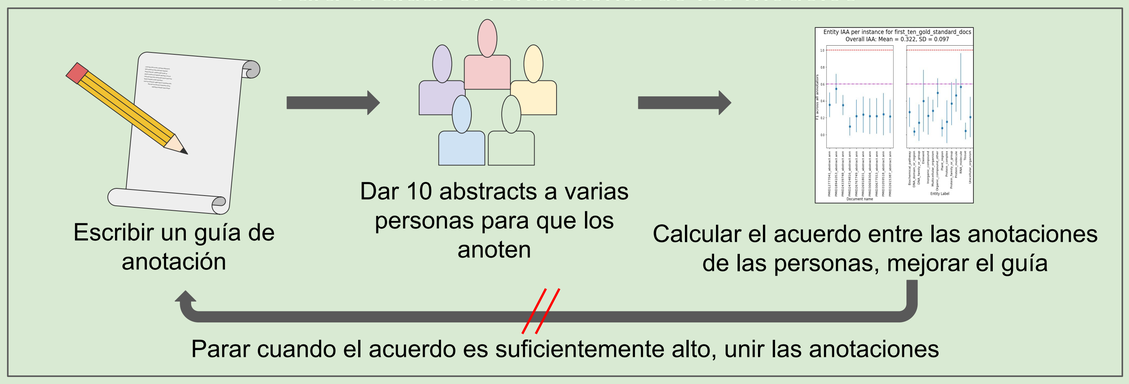
INTRODUCCIÓN

- Un grafo de conocimiento en el ámbito biológico consiste en objetos biológicos, como genes o proteínas, conectadas con relaciones entre ellos; esta información se extrae automáticamente de fuentes como artículos científicos
- Se puede usar en combinación con algoritmos diversos para predecir hipóteses científicos
- Los algoritmos que extraen automáticamente información de artículos científicos necesitan datos etiquetados con las palabras y relaciones que se deben extraer
- Por esa razón, este método apenas se ha usado en el ámbito de la biología vegetal molecular; no existe un conjunto etiquetado en este campo, y ningún trabajo ha documentado el uso de conjuntos de otros campos y su eficaz
- Este proyecto intenta averiguar el efecto de la especificad de los datos de entrenamiento en el rendimiento de los modelos, para mejor guiar los esfuerzos de etiquetar datos nuevos

METODOLOGÍA

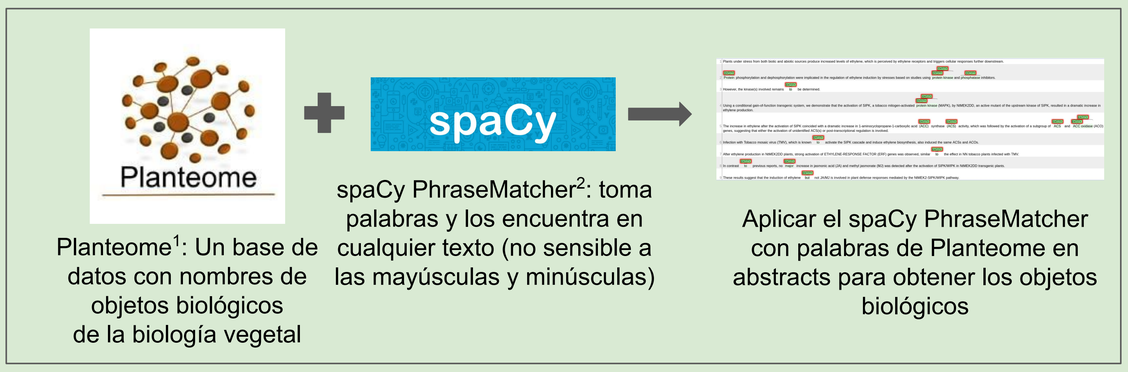
CONJUNTO ETIQUETADO EN LA BIOLOGÍA VEGETAL

Para evaluar el rendimiento de los modelos



BENCHMARKS NO NEURONALES

Un método sencillo basado en reglas, pero específica a la biología vegetal



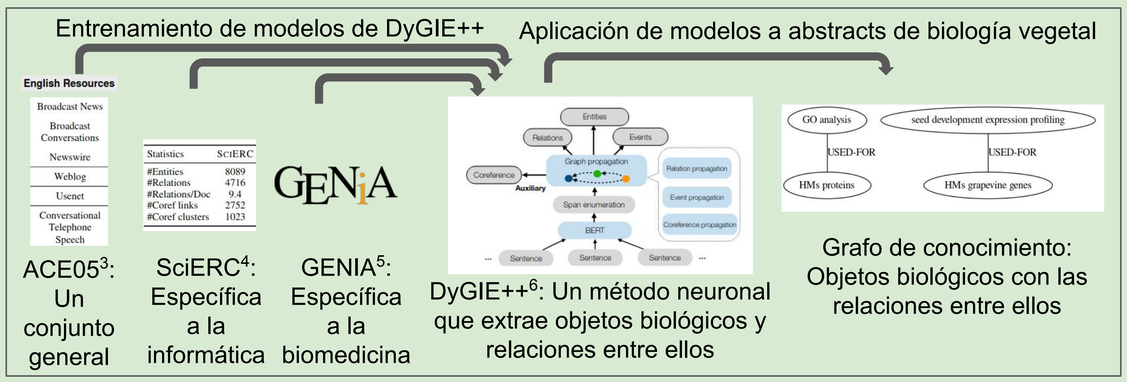
Planteome[1]: Un base de datos con nombres de objetos biológicos de la biología vegetal

spaCy PhraseMatcher[2]: toma palabras y los encuentra en cualquier texto (no sensible a las mayúsculas y minúsculas)

Aplicar el spaCy PhraseMatcher con palabras de Planteome en abstracts para obtener los objetos biológicos

METODOS NEURONALES

Métodos avanzados neuronales, pero sin la especificidad de la biología vegetal



ACE05[3]: Un conjunto general

SciERC[4]: Específica a la informática

GENIA[5]: Específica a la biomedicina

DyGIE++[6]: Un método neuronal que extrae objetos biológicos y relaciones entre ellos

Grafo de conocimiento: Objetos biológicos con las relaciones entre ellos

RESULTADOS

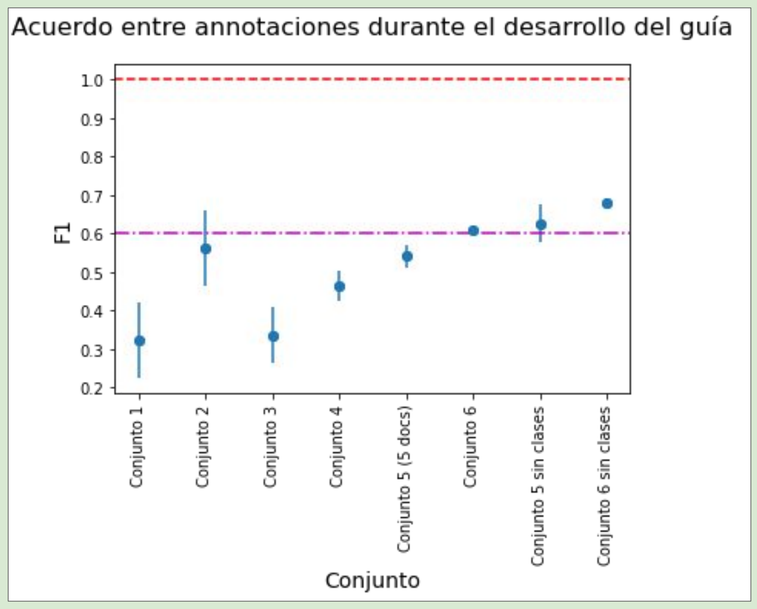


Figura 1. Acuerdo entre anotaciones durante el desarrollo del guía de anotación. Entre la anotación de cada uno de los conjuntos 1-6, retroalimentación de los anotadores se usó para mejorar el guía de anotación. La meta era lograr un F1 de 0.6 (la línea morada) antes de usar las anotaciones para evaluar el rendimiento de los modelos. Una puntuación perfecta es 1.0 (la línea roja). Los últimos dos números son los acuerdos de los conjuntos 5 y 6, pero calculado sin los clases (e.g. “proteína”, “gen”) que se dieron a las frases. Como los modelos tampoco dan clases a las frases, no es necesario considerarlas en la calculación de la puntuación.

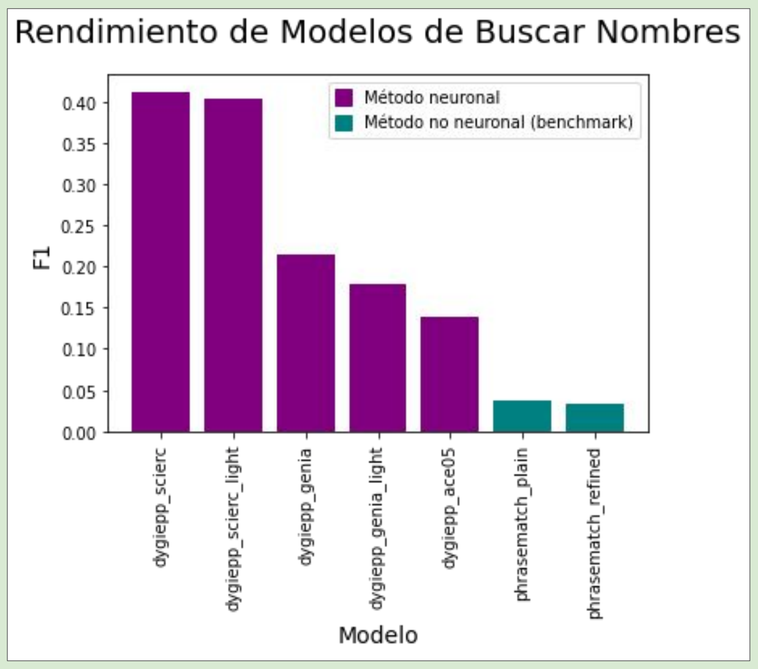


Figura 2. Rendimiento de los modelos. El conjunto 6 se usó para evaluar el rendimiento de los modelos. La puntuación máxima es un 1.0. Como se esperaba, los métodos no neuronales tienen un rendimiento mucho más bajo que los métodos neuronales. El modelo neuronal que se entrenó en ACE05, que es texto general, tiene un rendimiento muy bajo, y eso también se esperaba. Sin embargo, como el conjunto etiquetado de GENIA es de la biomedicina y por esta razón se considera más “cerca” al tema de la biologja vegetal que el conjunto de SciERC (informática), SciERC tiene un rendimiento mucho más alto que GENIA.

CONCLUSIONES

- Un método iterativo de mejorar el guía de anotación resulta en un acuerdo bastante alto de las anotaciones
- Asignar clases a las frases resulta un poco más difícil para los anotadores, que tienen un acuerdo más alto sin ellas
- Los modelos neuronales tienen un rendimiento mucho más alto que los benchmarks, aunque los benchmarks son específicas a la biología vegetal
- No obstante, el rendimiento de los modelos neuronales no necesariamente corresponde con la cercanía del tema de sus datos de entrenamiento al tema del texto en el que se aplica el modelo

BIBLIOGRAFÍA

1. J. Preece, J. Elser, y P. Jaiswal, «Planteome annotation wiki: a semantic application for the community curation of plant genotypes and phenotypes», en Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences - SWAT4LS ’11, London, United Kingdom, 2012, pp. 96-97. doi: 10.1145/2166896.2166921.
2. M. Honnibal y I. Montani, «spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing», 2017.
3. «The ACE 2005 (ACE05) Evaluation Plan», oct. 03, 2005. http://web.archive.org/web/20090902090933/http://www.itl.nist.gov/iad/mig//tests/ace/2005/doc/ace05-evalplan.v3.pdf (accedido nov. 08, 2021).
4. Y. Luan, L. He, M. Ostendorf, y H. Hajishirzi, «Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction», arXiv:1808.09602 [cs], ago. 2018, Accedido: oct. 26, 2020. [En línea]. Disponible en: http://arxiv.org/abs/1808.09602
5. Ohta, Tomoko, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. The GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the Proceedings of the Human Language Technology Conference (HLT 2002). San Diego, USA, March 2002.
6. D. Wadden, U. Wennberg, Y. Luan, y H. Hajishirzi, «Entity, Relation, and Event Extraction with Contextualized Span Representations», arXiv:1909.03546 [cs], sep. 2019, Accedido: mar. 15, 2021. [En línea]. Disponible en: http://arxiv.org/abs/1909.03546

AGRADECIMIENTOS

Quiero agradecer a todos los que me han ayudado con las anotaciones: Kenia Segura-Abá, Thilanka Ranaweera, Ally Schumacher, Melissa Lehti-Shiu, Abigail Seeger, y Brianna Brown, así como un miembro de mi comité asesor, Mohammad Ghassemi, por sus ideas y retroalimentación. También quiero agradecer a todos los que han desarrollado los programas y bases de datos que uso en este proyecto, especialmente por su buena documentación y sus respuestas rápidas.