The effect of the specificity of training data for knowledge graphs: A study in molecular plant biology

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

Knowledge graphs are a natural language processing technique that can be useful in automatically extracting information from scientific articles and in generating hypotheses. This project explores the effect of the specificity of training data for automated information extraction models in plant biology.

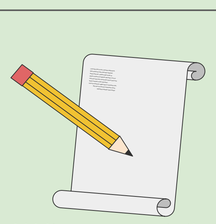
INTRODUCTION

- A knowledge graph in the biological sphere consists of biological objects, like genes or proteins, connected by the relations between them; this information is automatically extracted from sources like scientific articles
- They can be used in combination with a diverse set of algorithms to predict scientific hypotheses
- Automatic information extraction algorithms need datasets labeled with the words and relations that should be extracted
- For this reason, this method has barely been used in the sphere of molecular plant biology; there is no existing labeled corpus in this domain, and the use of corpora and their efficacy from other domains has also not been documented
- This project intends to determine the effect of the specificity of training data on the performance of these models, in order to better guide efforts to label new data

METHODS

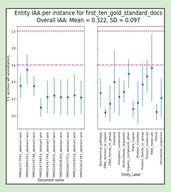
LABELED CORPUS IN PLANT BIOLOGY

In order to evaluate model performance



Write an annotation guideline

Give 10 abstracts to various individuals for annotation



Calculate the agreement between annotations, improve the guidelines

Stop when agreement is sufficiently high, unify the annotations

NON-NEURAL BENCHMARKS

A simple rule-based method, but specific to plant biology

Planteome[1]: A database with the names of biological objects in plant biologyl

spaCy PhraseMatcher[2]: takes words and finds them in any text (non-case sensitive)

Apply spaCy PhraseMatcher with words from Planteome on abstracts to obtain the biological objects

NEURAL METHODS

Advanced neural methods, but without specificity to plant biology

Train DyGIE++ models

Apply to plant biology abstracts

English Resources

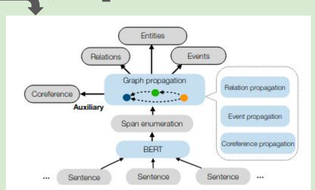
Broadcast News
Broadcast Conversations
Newswire
Weblog
Usenet
Conversational Telephone Speech

ACE05[3]: A general corpus

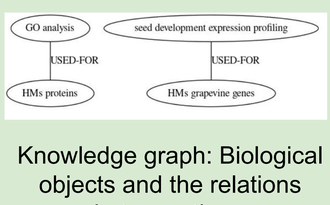
Statistics	SciERC
#Entities	8089
#Relations	4716
#Relations/Doc	9.4
#Coref links	2752
#Coref clusters	1023

SciERC[4]: Specific to computer science

GENIA[5]: Specific to biomedical sciences



DyGIE++[6]: A neural method that extracts biological objects and the relations between them



Knowledge graph: Biological objects and the relations between them

RESULTS

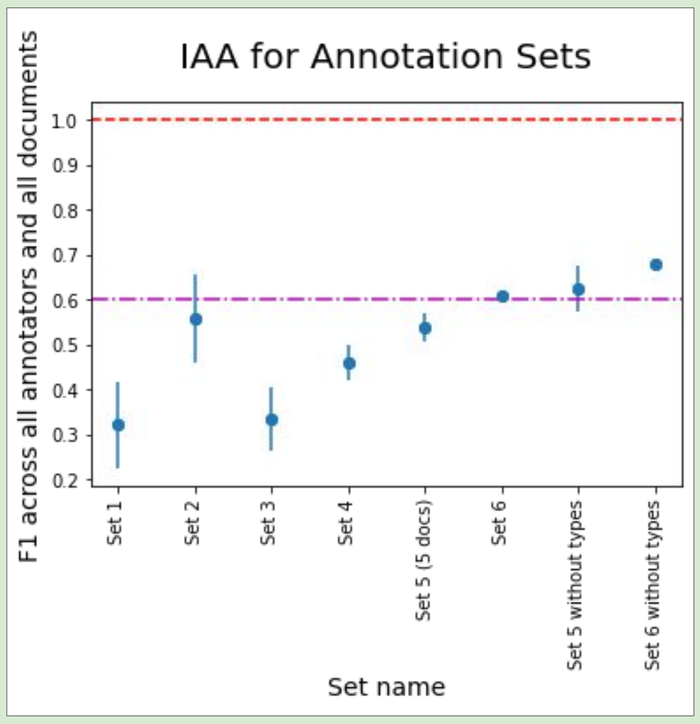


Figure 1. Agreement between annotations during the development of the annotation guidelines. Between the annotation of each of the 6 annotation sets, annotator feedback was used to improve the annotation guidelines. The goal was to achieve an F1 of 0.6 (the purple line) before using the annotations to evaluate model performance. A perfect score is 1.0 (the red line). The last two numbers are the agreements of the 5th and 6th annotation sets the without entity types (e.g. “protein”, “gene”) that are given to the spans. Since the downstream models also don't give types to the spans, it’s not necessary to consider them in the calculation of the score.

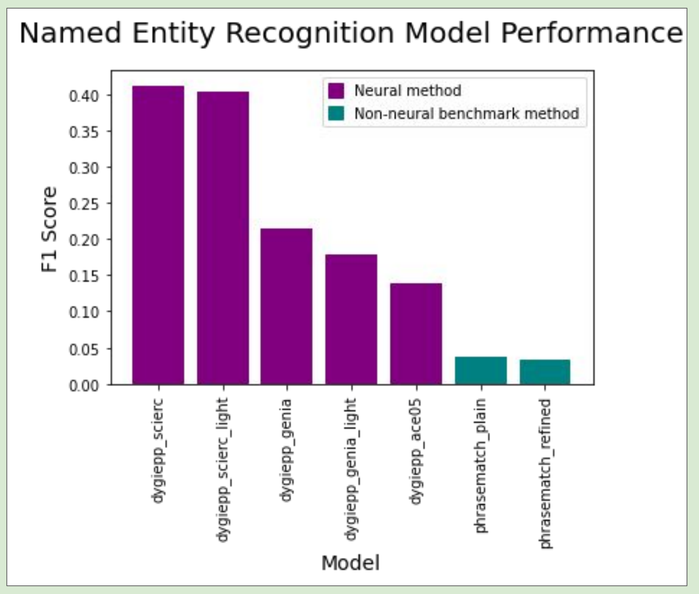


Figure 2. Model performance. Annotation set 6 was used to evaluate model performance. The maximum performance is 1.0. As expected, the non-neural methods have a much lower performance than the neural methods. The neural method trained on ACE05, which is a general corpus, has a very low performance, which was also expected. However, while the GENIA corpus is from the biomedical sciences and therefore can be considered “closer” to the topic of plant biology than the SciERC (computer science) corpus, SciERC has a performance that is much higher than that of GENIA.

CONCLUSIONS

- An iterative method for improving the annotation guidelines results in a fairly high agreement between annotations
- Assigning classes to the spans is somewhat difficult for the annotators, as they have a higher agreement without them
- The neural models have a much higher performance than the benchmarks, although the benchmarks are specific to plant biology
- However, the performance of the neural models doesn’t necessarily correspond with the relative “closeness” of their topics to the topic on which the models are being applied

BIBLIOGRAPHY

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. 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

ACKNOWLEDGMENTS

I would like to thank everyone who has helped me with annotations: Kenia Segura-Abá, Thilanka Ranaweera, Ally Schumacher, Melissa Lehti-Shiu, Abigail Seeger, and Brianna Brown, as well as my committee member Mohammad Ghassemi for his input and ideas for this project. I also want to thank all the developers of the programs and databases I use in this project, especially for their good documentation and quick responses.