Comparison of Natural Language Processing Tools for Automatic Gene Ontology Annotation of Scientific Literature

Lucas Beasley and Prashanti Manda

Department of Computer Science, University of North Carolina at Greensboro, NC, USA

***Abstract*—Manual curation of scientific literature for ontology-based knowledge representation has proven infeasible and unscalable to the large and growing volume of scientific literature. Automated annotation solutions that leverage text mining and Natural Language Processing (NLP) have been developed to ameliorate the problem of literature curation. These NLP approaches use parsing, syntactical, and lexical analysis of text to recognize and annotate pieces of text with ontology concepts. Here, we conduct a comparison of four state of the art NLP tools at the task of recognizing Gene Ontology concepts from biomedical literature using the Colorado Richly Annotated Full-Text (CRAFT) corpus as a gold standard reference. We demonstrate the use of semantic similarity metrics to compare NLP tool annotations to the gold standard.**

I. Introduction

There has been a rapid increase in the number of scientific articles published each year [1]. However, the majority of information in these scientific articles remains in the form of free text, and is therefore, opaque to computational analyses [2]. In several data intensive fields such as Biology, ontologies have been adopted as the de-facto mode of data representation to enable data integration, sharing, and, to make data computationally amenable [3].

While ontologies have helped transform information from free-text to a machine readable form, the process of this data transformation involves a huge bottleneck. The majority of ontology-based data annotation is performed via manual curation of scientific literature - the process of reading and annotating parts of text with one or more ontology concepts [4]. Manual curation is tedious, time consuming, and highly unscalable to the growing body of scientific literature.

To counter these difficulties, there has been a push to develop automated solutions based on Natural Language Processing (NLP) that can automatically read literature and identify ontology concepts from text, thereby performing automated annotation of literature.

One of the primary tasks for these NLP methods is Named Entity Recognition (NER) - identifying entities such as genes, proteins, or ontology concepts from text [5]. NER is an important component of information extraction and annotation for a wide range of domains such as biomedical research, biology, etc [6]. In other applications, NER is one of the crucial preliminary steps for subsequent creation of complex ontology-based expressions [7]. For example, the Entity Quality (EQ) annotation format is widely used to describe clinical and evolutionary phenotypes [8], [9]. Curators identify appropriate ontology concepts to represent the affected Entity and the Quality in a phenotype, and then combine the concepts to create an EQ expression. In more complex annotations, the Entity component might comprise multiple concepts that are combined using relational/spatial terms. In all of these scenarios, the first and crucial step is recognizing individual ontology concepts from text before building complex expressions.

Here, we focus on ontology-based Named Entity Recognition and conduct a formal comparison of methods and tools for recognizing ontology concepts from scientific literature in an automated manner. We present a comparison of four state of the art concept recognition tools (MetaMap [10], NCBO Annotator [1], Textpresso [11], and SciGraph [12]). We use the Colorado Richly Annotated Full-Text (CRAFT) corpus [13] as a Gold Standard reference to compare and assess the performance of these four NLP tools.

The CRAFT corpus contains 67 open access, full length biomedical articles annotated with concepts from several ontologies (such as Gene Ontology [3], Protein Ontology [14], Sequence Ontology [15], etc.). In our experiment, we employed each of the four NLP tools to annotate the articles in the CRAFT corpus with Gene Ontology concepts and subsequently compared the resulting annotations to CRAFT's GO annotations in order to assess the tool's annotation performance.

Precision and Recall are the most widely used metrics to assess the performance of any information retrieval system. However, these traditional metrics of performance don't take into account the notion of partial information retrieval. For example, in a traditional database, information is either retrieved or not retrieved by a search system, leading to a boolean characterization of performance. However, when aiming to "retrieve" an appropriate ontology concept for a piece of text, a tool might partially retrieve the concept as compared to a Gold Standard. For example, the Gold Standard might annotate a piece of text with the Gene Ontology concept "*apoptotic process*" while an NLP tool might annotate the same text with the concept "*programmed cell death*". From the perspective of Precision and Recall, this is an instance of information not being retrieved or incorrect retrieval. However, subsumption reasoning within the Gene Ontology indicates that the two concepts are closely related since "*apoptotic*

p_manda@uncg.edu

process" is a direct sub-class of "*programmed cell death*". Thus it can be argued the NLP tool retrieves a semantically similar, albeit, a slightly more general version of the Gold Standard's annotation.

We propose the use of semantic similarity metrics to capture the degree of relatedness between an NLP tool's performance as compared to the Gold Standard to account for the notion of partial retrieval for ontology-based systems. Semantic similarity is defined as the degree of relatedness between ontology concepts or objects based on their ontology annotations [16]. The use of other metrics that utilize the ontology hierarchy to evaluate NLP annotations can also be seen in BioCreative IV [17].

II. Related Work

There have been several studies that provide an overview of text mining and NLP techniques as applied to biological/biomedical literature [18–21].

One of the recent studies focusing on comparison of annotation tools by Funk et al. [22] conducts a comparison of three tools - MetaMap, NCBO Annotator, and ConceptMapper [23]. Performance of these tools is evaluated on eight biomedical ontologies using the CRAFT corpus as a gold standard. The paper explores different parameter setting for the tools to identify the most optimal combinations.

Most notable among these studies is from Kelley et al. [24], [25] as they investigate strategies to build the National Center for Biomedical Ontology Annotator (NCBO Annotator). Kelley et al. [25] compared two concept recognition tools, Mgrep [26] and MetaMap [10] using multiple data sets and vocabularies/ontologies of different sizes to evaluate the scalability and performance of each tool. Their results show that Mgrep consistently recognized fewer unique concepts than MetaMap; which led to a higher precision score in every evaluation with the exception of one. Kelley et al. use Precision and Recall to quantify the performance of the two tools therefore missing closely related but not exactly matching annotations retrieved by the tools. In contrast, we use semantic similarity metrics that are capable of quantifying exact matches and matches of varying degrees of relatedness to provide a more accurate assessment of a tool's annotation performance.

Similarly, Galeota and Pelizzola. [27] performed a comparison of MetaMap [10] and ConceptMapper. In their comparison they used a public repository (Sequence Read Archive) of raw experiment data and its corresponding metadata, three ontologies for annotating, and knowledge-based semantic similarity measures such as 'path finding' (shortest path between concepts) and Information Content based metrics. Their findings show that ConceptMapper outperformed MetaMap in precision and recall, however MetaMap performed better than ConceptMapper when presented with entire sentences.

III. Methods

Articles from the CRAFT corpus v2.0 were used as input for each NLP tool. The CRAFT corpus [28] contains 67 publicly available articles that have been manually annotated on six ontologies including the Gene Ontology.

The annotation performance of four tools - MetaMap, NCBO Annotator, SciGraph Annotator, and Textpresso was evaluated at the task of annotating the 67 articles in the CRAFT corpus with GO concepts. Below, we provide methodological details, execution configurations, and parameters of the four NLP tools.

A. NLP Tools

1) MetaMap: MetaMap is a program designed to recognize and annotate biomedical text to concepts in the Unified Medical Language System (UMLS) Metathesaurus [10]. The program operates via four primary steps. First, text is parsed to tag parts of speech such as nouns to provide a syntactic analysis. Parsed phrases are then used to generate variants that account for acronyms, abbreviations, synonyms, etc. Next, candidate strings/concepts within the UMLS Metathesaurus that match one or more of the above variants are identified to build a candidate set of annotations. The candidate set is further evaluated against the input text using several metrics to assess the strength of the annotations. The highest scoring candidate is used to form the final mapping for the phrase.

MetaMap was run using the scheduled batch processor with the following parameters:

- **-V USAbase**: Selects the USA base data version
- **-L 17**: Selects the 2017 version of the SPECIALIST lexicon
- **-Z 1718**: Selects the 2017AB UMLS Metathesaurus as the knowledge source
- **-E**: Indicates the end of a citation with the 'EOT' flag
- **-T**: Tagger output - lines up the tagged output on the lines below the input.
- **-I**: Displays the UMLS IDs for each tagged concept
- **-R GO**: Restricts annotations to the Gene Ontology

UMLS concepts identified by MetaMap were mapped to GO identifiers (GO release date 04/28/2017). In some instances, MetaMap generates multiple annotations corresponding to a piece of text (phrase). To determine the precise piece of text being annotated to a concept, we verified if the concept name of the annotation was a sub-string of the phrase. If yes, we recorded the annotation as being tagged to the sub-string and not to the larger phrase. If not, the annotation was associated to the larger phrase. For example, if the phrase "several thousand liver gene expression quantitative trait loci" was annotated with the GO concept "*gene expression*", the annotation would be tagged to "gene expression" in text because it is a sub-string of the original phrase. However, if the phrase "on the regulation of many physiological traits" was annotated to GO concept "*regulation of biological process*" then the annotation would be mapped to the full phrase "on the regulation of many physiological traits" rather than a specific sub-string within the phrase.

2) NCBO Annotator: The National Center for Biomedical Ontology (NCBO) Annotator [1] was developed to recognize ontology concepts from biological and biomedical text. NCBO Annotator is implemented to provide annotations to 207 different ontologies.

First, the text to be annotated is provided as input along with a dictionary containing concepts from one or more desired

ontologies. An ontology concept recognizer named Mgrep ([26]) is used to match input text to concepts in the dictionary. These matches called direct annotations are later expanded to create semantic expansions using ontology semantics and relationships.

For example, the is_a relationship can be used to identify subsumer concepts of direct annotations. The ontology concept recognizer also uses one-to-one cross mappings between ontology terms to identify new annotations. For example, direct annotations to a concept in one ontology can be expanded to gather new annotations from another ontology based on pre-existing bridge mappings between the two ontologies. NCBO also proposes the use of semantic similarity metrics to identify related concepts for direct annotations to create new annotations.

Using their REST API, NCBO was run with the following parameters on the Gene Ontology (release date 10/25/2017).

- **ontologies=GO**: Restricts the annotator to the Gene Ontology
- **text=**: Passes the UTF-8 encoded text to the annotator

3) Textpresso: Textpresso [11] uses the Textpresso ontology that contains concepts from 33 parent ontologies such as gene, cellular component, nucleic acid, organism, phenotype, drugs including the Gene Ontology. These parent categories may be further classified into one or more sub-categories.

Textpresso breaks free text into sentences and words and uses a text-to-XML converter that marks the text through and generates XML documents [11]. Users can opt to retrieve annotations from one or more parent/sub-categories along with searching specific portions of the text such as the abstract/body, etc. Textpresso was run using the default command with no additional parameters specified. The Textpresso ontology (release date 07/2011) at the time of execution used Gene Ontology version GO v1.934.

4) SciGraph Annotator: The Monarch annotation service [12] provides the SciGraph annotator that annotates user provided free text with ontology concepts and biological entities. SciGraph annotator marks the text with concepts from the Monarch knowledge graph that includes gene ontology terms, genes, diseases, and phenotypes. SciGraph also allows the user the flexibility of selecting any desired ontology for annotation. SciGraph was run with the **ann** parameter on the Monarch knowledge graph (as of 04/09/2018) that calls the annotation service.

B. Evaluation of Tool Annotation Performance

Annotations generated by the four NLP tools were compared to the CRAFT corpus for evaluating their annotation performance. Annotations generated by the four tools and those in the CRAFT corpus were represented in a consistent location-based format that indicated annotations corresponding to the piece of text between a starting character index and an ending character index. This representation enabled the comparison of annotation output across sources in a consistent manner.

Consider a piece of text with starting character index i and ending index j. We denote a CRAFT annotation corresponding to this text as $C_{i,j}$ and a tool annotation for the same text as $N_{i,j}$ where N represents one of the four tools. NLP tool annotations might be compared to those from CRAFT in one of the following cases:

- Exact Match: $C_{i,j} \neq \emptyset$, $N_{i,j} \neq \emptyset$, and $C_{i,j} = N_{i,j}$
  An exact annotation match occurs when a piece of text is annotated with the same ontology concept in CRAFT and a tool. For example, both the tool and CRAFT corpus annotate the text "development" (i = 191, j = 202) with the GO concept "*developmental process*" (GO:0032502)
- Partial Match: $C_{i,j} \neq \emptyset$ and $N_{i,j} \neq \emptyset$, $C_{i,j} \neq N_{i,j}$
  A partial annotation match occurs when a piece of text is annotated both by CRAFT and a tool but the ontology concepts used vary between the two sources. For example, both the tool and CRAFT corpus annotate the text "antibody" (i = 5964, j = 5972). However, CRAFT used the GO concept *"immunoglobulin complex"* (GO:0019814) while the tool used *"B cell receptor complex"* (GO:0019815).
- False Positive: $C_{i,j} = \emptyset$, $N_{i,j} \neq \emptyset$
  A false positive annotation occurs when a tool generates an annotation for a piece of text that does not contain an annotation in CRAFT. For example, a tool annotated "homeostasis" (i = 233, j = 244) with "*homeostatic process*" (GO:0042592), but there was no annotation in CRAFT for the same text.
- False Negative: $C_{i,j} \neq \emptyset$, $N_{i,j} = \emptyset$
  A false negative annotation occurs when a tool fails to annotate a piece of text that has been annotated in CRAFT. For example, the CRAFT corpus annotates the text "immune function" (i = 204, j = 219) with "*immune system process*" (GO:0002376), but the tool did not annotate that piece of text.

C. Semantic Similarity

The accuracy of a tool's annotation matches (exact and partial) in comparison to CRAFT was assessed using Jaccard semantic similarity. The Jaccard similarity (J) of two ontology concepts/classes (A, B) in an ontology is defined as the ratio of the number of classes in the intersection of their subsumers over the number of classes in their union of their subsumers [16], [29].

$$J(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

where $S(A)$ is the set of classes that subsume A. Jaccard similarity ranges from 0 (no similarity) to 1 (exact match). Jaccard similarity is designed to measure similarity between ontology concepts based on their proximity to each other in the ontology - the farther two concepts are from each other, the less similar they are.

IV. Results and Discussion

The CRAFT corpus contains 67 open access, full-length papers annotated with concepts from 6 ontologies. The corpus contains 91,446 total annotations and 3,242 unique annotations spanning 6 ontologies. Table I shows the total and unique number of annotations categorized by ontology. CRAFT contains a

total of 32,416 GO annotations (35.4% of all ontology annotations) with an average of 416 annotations per paper. These GO annotations are distributed across three sub-ontologies Cellular Component (CC), Molecular Function (MF), and Biological Process (BP) (Figure 1). Results in Figure 1 indicate that a large portion of GO annotations in the CRAFT belong to the Biological process sub-ontology with Molecular Function accounting for the least number of annotations.

TABLE I. DISTRIBUTION OF ANNOTATIONS IN THE CRAFT CORPUS BY ONTOLOGY

Ontology	Number of total annotations	Number of unique annotations
Gene Ontology	32,416	1,236
Sequence Ontology	22,090	260
Protein Ontology	15,594	889
Chemical Entities of Biological Interest	8,137	553
NCBI Taxonomy	7,449	149
Cell Type Ontology	5,760	155

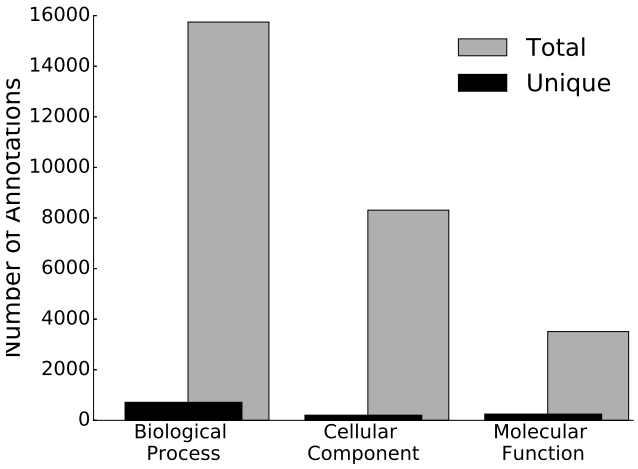

Fig. 1. Distribution of annotations across the three sub-ontologies of the Gene Ontology

Four NLP tools (MetaMap, NCBO Annotator, Textpresso, and SciGraph) were used to annotate the 67 CRAFT articles with GO concepts. First, we observed the performance of the four tools using the total number of unique and non-unique annotations generated by each tool as compared to the CRAFT Corpus (Table II). Here, we see that Textpresso retrieves the most number of annotations among the four tools, surprisingly, about 24% more than the CRAFT corpus. However, Textpresso annotations contain a very low number of unique GO terms indicating that these annotations do not span across the Gene Ontology well. MetaMap retrieves about 82% of the CRAFT total annotation count and 94% of CRAFT's unique annotation count. NCBO Annotator and SciGraph generate a substantially lower number of annotations as compared to MetaMap and Textpresso.

We also examined if the distributions of NLP annotations at different depths of the GO were similar to CRAFT annotations. We observe striking similarities in the distributions of annotations across GO depths for CRAFT and the NLP tools except Textpresso (Figure 2). The majority of Textpresso annotations belong at level 1 (direct children of the root) with a distinct lack of annotations at deeper levels. This trend is in stark contrast to trends seen with CRAFT and the other three tools where the number of unique terms increases with increasing depth peaking at level 6 and declining after. This result is unsurprising given the low number of unique GO terms retrieved by Textpresso (Table II).

TABLE II. NUMBER OF TOTAL AND UNIQUE GO ANNOTATIONS IN THE CRAFT CORPUS AND RETRIEVED BY EACH TOOL.

Annotation Source	Total GO annotations (% CRAFT)	Unique GO annotations (% CRAFT)
CRAFT	32,416	1,236
Textpresso	40,509 (124 %)	37 (3%)
MetaMap	26,815 (82%)	1,174 (94%)
NCBO Annotator	15,611 (48%)	894 (72%)
SciGraph Annotator	13,149 (40%)	780 (63%)

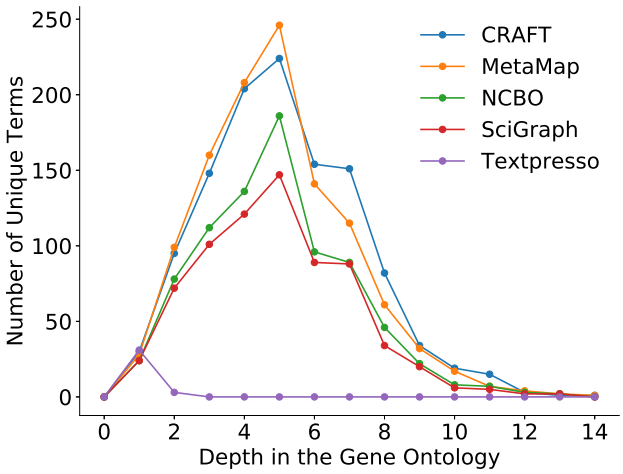

Fig. 2. Comparison of the distribution of annotations at different levels of the Gene Ontology from different annotation sources.

Next, we examined each tool's performance with respect to the proportion of false positives and negatives (Table III). These results indicate that a staggering proportion of annotations generated by MetaMap and Textpresso, the two tools generating the most annotations (Table II) are false positives or false negatives. In comparison, SciGraph and NCBO Annotator generate lower false positives.

We hypothesize that the large number of false positives and false negatives from MetaMap might be an artifact of our interpretation of MetaMap's results. Without a clear mapping between the exact piece of text being annotated to a concept, we used a fuzzy matching between the tagged phrase and the annotation which might have contributed to these false positives.

We suspect that a proportion of false positives generated by the tools might be because CRAFT annotations were generated

using an outdated version of the GO whereas the tools use newer versions of the GO with a more complete set of ontology concepts.

TABLE III. FALSE POSITIVE AND FALSE NEGATIVE ANNOTATIONS BY THE NLP TOOLS AS COMPARED TO THE CRAFT CORPUS.

Annotation Source	False Positives (% annotations)	False negatives (% annotations)
MetaMap	25,823 (96.3%)	31,446 (97.0%)
Textpresso	26,692 (65.9%)	22,937 (70.8%)
SciGraph Annotator	6,494 (49.4%)	27,607 (85.2%)
NCBO Annotator	6,914 (44.3%)	24,892 (76.8%)

Next, we focused on exact and partial matches between the tools and CRAFT. Surprisingly, we notice that most of the tools generate appreciably more exact matches as compared to partial matches (Table IV). These numbers also indicate that a rather small proportion (2-18%) of total annotations generated by a tool are exact or partial matches to CRAFT. We see that NCBO Annotator produces the highest number of exact matches followed by SciGraph. Surprisingly, despite it's high number of generated annotations, MetaMap produces just 2.7% exact matches and 0.4% partial matches.

TABLE IV. NUMBER OF EXACT AND PARTIAL MATCHES BETWEEN GO ANNOTATIONS IN THE CRAFT CORPUS AND THOSE GENERATED BY EACH TOOL.

Annotation Source	Number of exact matches (% matches)	Number of partial matches (% matches)
NCBO Annotator	5,921 (18.3%)	2,776 (8.6%)
SciGraph Annotator	5,016 (15.5%)	1,639 (5.1%)
Textpresso	4,006 (12.4%)	9,811 (30.3%)
MetaMap	866 (2.7%)	126 (0.4%)

The four tools generated a total of 15,809 exact matches and 14,352 partial matches to CRAFT. We analyzed the overall quality and accuracy of these matched annotations by comparing them to corresponding CRAFT annotations using Jaccard semantic similarity. Here, we see that MetaMap has a very high similarity to CRAFT (90%) followed by SciGraph and NCBO. This result further reinforces our hypothesis that the MetaMap false positives are likely a result of our fuzzy interpretation of its annotation output. It is important to note that although MetaMap results in the highest semantic similarity score, the pool of annotations being analyzed is substantially smaller as compared to the other tools. Textpresso annotations show the lowest similarity scores. Given the discrepancy in the depth of distribution between Textpresso's annotations and the CRAFT corpus, it is unsurprising to see the low semantic similarity score between Textpresso and CRAFT. As in the above tests, SciGraph and NCBO perform well (75-81% similarity) with relatively few false positives and false negatives.

Another factor that could affect the semantic similarity scoring and the performance of the tools is the difference in ontology versions used for annotation between CRAFT and the various tools. Some of the NLP tools compared in this study do not support user specified ontology versions for annotation.

We examined if it is easier for NLP tools to annotate GO concepts with simple names (as measured by number of words in the concept name) as compared to concepts with multiple word names. Since the primary process employed by the NLP tools here is to analyze and match text lexically to concept names, we hypothesized that it might be easier to recognize short concept names as compared to longer ones. To explore this, we plotted the semantic similarity of annotation matches categorized by the number of words in the concept names. Interestingly, we do not see any consistent decreases in semantic similarity as the number of words in the GO concept name increases (Figure 4). This indicates that the ease of annotation does not decline as ontology concept names get longer and more complex.

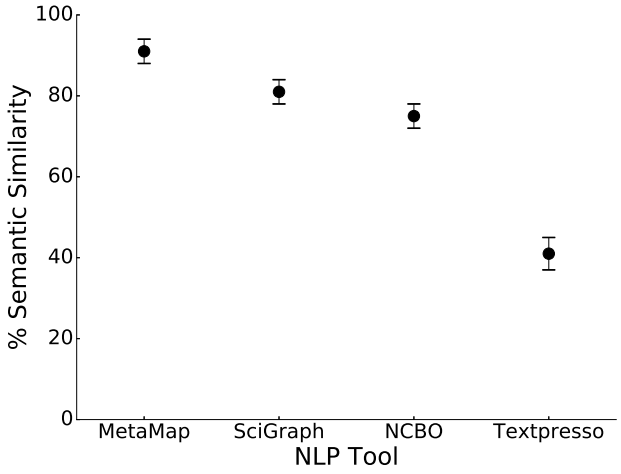

Fig. 3. Semantic similarity comparison of partial and exact matches between NLP tool annotation and the CRAFT corpus. Error bars indicate two standard errors of the mean.

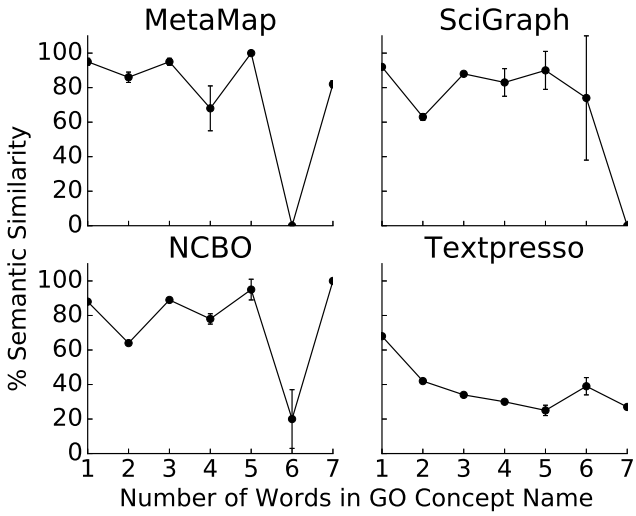

Fig. 4. Comparison of annotation performance indicated by semantic similarity for GO concept names of different lengths. Error bars represent two standard errors of the mean.

V. Conclusions

Here, we conducted a comparison of MetaMap, NCBO Annotator, Textpresso, and SciGraph at the task of annotating scientific literature with Gene Ontology concepts. We found that while Textpresso generated the most amount of annotations, a large proportion of the annotations were false positives or false negatives. Also, Textpresso annotations featured a disproportionate number of high level GO terms as compared to CRAFT and other tools. MetaMap generated the largest number of false positives and false negatives but had the highest similarity to CRAFT based on partial and exact matches. Although NCBO and Scigraph retrieve lower annotations as compared to the other tools, their performance in terms of false positives and false negatives is better in comparison to the other tools. The semantic similarity of NCBO and SciGraph is very comparable to the top performer, MetaMap. As compared to MetaMap and Textpresso which show extreme results, SciGraph and NCBO perform well on all evaluation categories indicating potential utility for scientific applications. Overall, the proportion of false positives and false negatives across all the tools indicates that there is substantial room for improvement in NLP tools for ontology-based Named Entity Recognition.

VI. Data and Software

The CRAFT data source can be found at http://bionlp-corpora.sourceforge.net/CRAFT/. Code for results and analysis can be found at https://github.com/lucasbeasley/average-jaccard.

VII. Acknowledgments

This work was supported by the University of North Carolina at Greensboro Giant Steps initiative funding to Manda. The authors acknowledge valuable technical assistance from MetaMap's development team.

References

[1] C. Jonquet, N. Shah, C. H Youn, M. Musen, C. Callendar, and M.-A. Storey, "Ncbo annotator: Semantic annotation of biomedical data," 01 2009.

[2] J. P. Balhoff, W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. Mabee, P. E. Midford, M. Westerfield, and T. J. Vision, "Phenex: ontological annotation of phenotypic diversity," *PLoS One*, vol. 5, no. 5, p. e10500, 2010.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[4] W. Dahdul, T. A. Dececchi, N. Ibrahim, H. Lapp, and P. Mabee, "Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy," *Database*, vol. 2015, 2015.

[5] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, "Text mining and ontologies in biomedicine: making sense of raw text," *Briefings in bioinformatics*, vol. 6, no. 3, pp. 239–251, 2005.

[6] U. Yasavur, R. Amini, C. L. Lisetti, and N. Rishe, "Ontology-based named entity recognizer for behavioral health." in *FLAIRS Conference*, 2013.

[7] H. Cui, W. Dahdul, A. T. Dececchi, N. Ibrahim, P. Mabee, J. P. Balhoff, and H. Gopalakrishnan, "Charaparser+ eq: Performance evaluation without gold standard," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–10, 2015.

[8] G. V. Gkoutos, C. Mungall, S. Dolken, M. Ashburner, S. Lewis, J. Hancock, P. Schofield, S. Kohler, and P. N. Robinson, "Entity/quality-based logical definitions for the human skeletal phenome using pato," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 7069–7072.

[9] P. Manda, J. P. Balhoff, H. Lapp, P. Mabee, and T. J. Vision, "Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution," *genesis*, vol. 53, no. 8, pp. 561–571, 2015.

[10] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[11] H.-M. Mller, E. E. Kenny, and P. W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," *PLOS Biology*, vol. 2, no. 11, 09 2004. [Online]. Available: https://doi.org/10.1371/journal.pbio.0020309

[12] C. J. Mungall, J. A. McMurry, S. Köhler, J. P. Balhoff, C. Borromeo, M. Brush, S. Carbon, T. Conlin, N. Dunn, M. Engelstad *et al.*, "The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic acids research*, vol. 45, no. D1, pp. D712–D722, 2016.

[13] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake *et al.*, "Concept annotation in the craft corpus," *BMC bioinformatics*, vol. 13, no. 1, p. 161, 2012.

[14] D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, M. Caudy, H. J. Drabkin, P. Deustachio, A. V. Evsikov, H. Huang *et al.*, "The protein ontology: a structured representation of protein forms and complexes," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D539–D545, 2010.

[15] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The sequence ontology: a tool for the unification of genome annotations," *Genome biology*, vol. 6, no. 5, p. R44, 2005.

[16] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS computational biology*, vol. 5, no. 7, p. e1000443, 2009.

[17] Y. Mao, K. Van Auken, D. Li, C. N. Arighi, P. McQuilton, G. T. Hayman, S. Tweedie, M. L. Schaeffer, S. J. Laulederkind, S.-J. Wang *et al.*, "Overview of the gene ontology task at biocreative iv," *Database*, vol. 2014, 2014.

[18] N. Tiffin, J. F. Kelso, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide, "Integration of text-and data-mining using ontologies successfully selects disease gene candidates," *Nucleic acids research*, vol. 33, no. 5, pp. 1544–1552, 2005.

[19] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[20] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia, "Evaluation of text-mining systems for biology: overview of the second biocreative community challenge," *Genome biology*, vol. 9, no. 2, p. S1, 2008.

[21] M. Krallinger, R. A.-A. Erhardt, and A. Valencia, "Text-mining approaches in molecular biology and biomedicine," *Drug discovery today*, vol. 10, no. 6, pp. 439–445, 2005.

[22] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, and K. Verspoor, "Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters," *BMC bioinformatics*, vol. 15, no. 1, p. 59, 2014.

[23] M. A. Tanenblatt, A. Coden, and I. L. Sominsky, "The conceptmapper approach to named entity recognition." in *LREC*, 2010, pp. 546–51.

[24] N. Bhatia, N. H. Shah, D. L. Rubin, A. P. Chiang, and M. Musen, "Comparing concept recognizers for ontology-based indexing: Mgrep vs. metamap," *Paper accepted to the AMIA Summit on Translational Bioinformatics, San Francisco*, 2009.

[25] N. H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang, and M. A. Musen, "Comparison of concept recognizers for

building the open biomedical annotator," *BMC Bioinformatics*, vol. 10, no. 9, p. S14, Sep 2009. [Online]. Available: https://doi.org/10.1186/1471-2105-10-S9-S14

[26] M. Dai, N. Shah, W. Xuan *et al.*, "An efficient solution for mapping free text to ontology terms. amia summit on translational bioinformatics," *San Francisco CA*, 2008.

[27] E. Galeota and M. Pelizzola, "Ontology-based annotations and semantic relations in large-scale (epi)genomics data," *Briefings in Bioinformatics*, vol. 18, no. 3, pp. 403–412, 2017. [Online]. Available: http://dx.doi.org/10.1093/bib/bbw036

[28] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, and L. E. Hunter, "Concept annotation in the craft corpus," *BMC Bioinformatics*, vol. 13, no. 1, p. 161, Jul 2012. [Online]. Available: https://doi.org/10.1186/1471-2105-13-161

[29] M. Mistry and P. Pavlidis, "Gene ontology term overlap as a measure of gene functional similarity," *BMC bioinformatics*, vol. 9, no. 1, p. 327, 2008.