How to measure the consistency of the tagging of scientific papers?

Anonymous authors

Abstract

A collection of scientific papers is often accompanied by tags: keywords, topics, concepts etc., associated with each paper. Sometimes these tags are human-generated, sometimes they are machine-generated. We propose a simple measure of the consistency of the tagging of scientific papers: whether these tags are predictive for the citation graph links. Since the authors tend to cite papers about the topics close to those of their publications, a consistent tagging system could predict citations. We present an algorithm to calculate consistency, and experiments with human- and machine-generated tags. We show that augmentation, i.e. the combination of the manual tags with the machine-generated ones, can enhance the consistency of the tags. We further introduce cross-consistency, the ability to predict citation links between papers tagged by different taggers, e.g. manually and by a machine. Cross-consistency can be used to evaluate the tagging quality when the amount of labeled data is limited.

1. Introduction

A part of a construction of a knowledge graph is the analysis of publications and adding to them tags: concept names, keywords, etc. This often involves natural language processing or other machine learning methods [Murty et al., 2018]. To develop such methods one must have a measure of success: one should be able to determine whether the given tagging is “good” or “bad”. The most direct way to test the machine produced tags is to compare them to the tags produced by humans. One creates a “golden set” of papers tagged by humans, and penalizes the algorithms for any deviation from these tags. There are, however, certain problems with this approach. First, human tagging is expensive—even more so for scientific papers, where human taggers must have a specialized training just to understand what the papers are about. Second, even the best human taggers' results are inconsistent. This provides a natural limitation for this method [Manning, 2011]. The latter problem is exacerbated when the tagging dictionary is large. For example, the popular US National Library of Medicine database of Medical Subject Headings (MeSH, https://www.nlm.nih.gov/mesh/) has just under 30 000 entries. A superset of

MeSH, Unified Medical Language System (UMLS, https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html) contains a staggering amount of 3 822 832 distinct concepts. It is doubtful a human can do a good job choosing the right tags from a dictionary so large. A domain expert usually deals with a subsystem of the dictionary, covering her area of expertise. This presents obvious difficulties for tagging papers about multidisciplinary research, that may require a combination of the efforts of several highly qualified taggers. Another problem is the evaluation of *tag augmentation*. Suppose we have papers tagged by humans, and we want to add machine-generated tags, for example, to improve the search function in the collection. Do the new tags actually add to the quality or subtract from it? How can we evaluate the result if our tags are by definition different from those produced by humans?

Thus a measure of the tagging quality other than a direct comparison with manually produced tags may be useful for the assessing the work of the tagging engines. This is especially important for an ongoing quality control of an engine that continuously ingests and tags fresh publications. In this paper we propose such a measure.

The idea for this measure is inspired by the works on graph embeddings [Hamilton et al., 2018, Grover and Leskovec, 2016]. In these works one tags graph nodes and compares different sets of tags. The usual comparison criterion is whether the tags can predict graph edges: nodes connected by an edge should have similar tags, while nodes not connected by an edge should have dissimilar tags. To use this approach we need to represent papers as nodes on a graph. A natural choice is the citation graph: and edge from paper A to paper B means that paper A cites paper B. This leads to the following assumptions:

1. Scientific papers cited by the given paper A are more similar to A than the other (non cited) papers.

2. A good tagging system must reflect this.

In other words, a good set of tags must be able to predict links on the citation graph, and the quality of the prediction reflects the quality of the tags. We will call this property *consistency*: a good tagger consistently gives similar tags to similar papers.

It is worth stressing that consistency is just one component of the quality of a tagger. If a tagger consistently uses keyword *library* instead of keyword *bread* [Borges, 1944], this measure would give it high marks, despite tags being obviously wrong. A way to overcome this deficiency is to calculate *cross-consistency* with a known “good” tagger. For example, we can tag some papers manually, and some papers using machine generated tags, and then predict citation links between these papers. This cross-consistency measures the similarity between these taggers. This application is interesting because it allows us to expand the number of labeled papers for evaluation of machine-based taggers. We can create a golden set of manually tagged papers, and then generate tags for the papers in their reference lists, and the random samples using the machine-based tagger. Since a typical paper cites many publications, this approach significantly expands the quantity of data available for training and testing.

To create a measure based on these ideas one should note that citation links strongly depend on the time the candidate for citation was published. Even a very relevant paper may not be cited if it is too old or too new. In the first case the citing authors may prefer

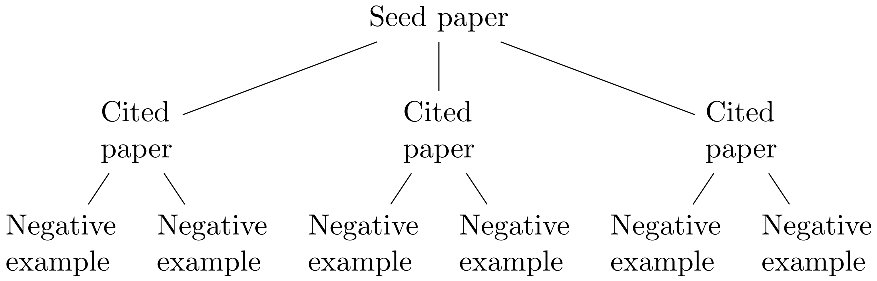

Figure 1: A seed paper, its cited papers and negative examples (here $k = 3$, $m = 2$).

a newer paper on the same topic. In the second case they may overlook the most recent publications. Therefore we recast our assumptions in the following way:

> A consistent tagging system should be able to predict citation links from a given paper to a set of simultaneously published papers.

The rest of the paper is organized as follows. In Section 2 we discuss the algorithm to calculate the consistency of the given tagging system. Experiments with this measure are discussed in Section 3. In Section 4 we present the conclusions.

2. Algorithm

The algorithm to calculate the consistency measure is shown as Algorithm 1. We select n seed papers. For each seed paper we take up to k random papers from its reference list, and label them with the label $y = 1$. These are our *cited papers*, or positive samples. For each cited paper we randomly choose k papers from the corpus with the same publication date as the cited paper. These are our *negative samples* (Figure 1). The notion of "publication date" depends on the granularity g of publication dates. We label the chosen negative samples with the label $y = 0$. Now we have a vector of labels y_i of size $k + km$, with k ones and km zeros. We calculate tags for the seed paper, positive and negative samples, and the number of overlapping tags between the seed paper and each of positive and negative samples. This gives us $k + km$ overlap numbers t_i of each seed paper. We can solve the one dimension classification problem $\mathbf{y} \sim \mathbf{t}$ and calculate its ROC curve [Fawcett, 2006]. The area under curve (AUC) for this problem is the measure of the consistency of the tagging. The average AUC and its variation are the numbers we are interested in.

Algorithm 1 can be used for calculation of consistency or cross-consistency of taggers. In the latter case we just choose different sources of tags for seed papers and samples.

3. Experiments

For experiments we used the papers extracted from the PubMed database (`https://www.ncbi.nlm.nih.gov/pubmed/`). These papers have MeSH tags attached manually by the human taggers. We added to them additional tags by processing papers' titles and abstracts using several packages. Gene names were identified using GNAT [Hakenberg et al., 2011], diseases were identified using DNORM [Leaman et al., 2013], and additional UMLS concepts

Algorithm 1 Calculation of consistency measure for the given tagging system

```
Input: Parameters: $n$ seed papers, $k$ cited papers per seed paper, $m$ negative samples per cited paper, granularity $g$ of the publication date
for all seed papers $\mathfrak{s}$ do
    Select $\min(k, \text{bibliography length})$ random papers from the reference list of $\mathfrak{s}$ (cited papers or positive samples). Label them as $y_i = 1$
    for all cited papers $\mathfrak{c}$ do
        Select $m$ random papers with same publication date as $\mathfrak{c}$, using date granularity $g$ (negative samples). Label them as $y_i = 0$.
    end for
    Tag the seed paper, positive and negative samples.
    for all positive and negative samples $\mathfrak{p}$ do
        Calculate the number $t_i$ of overlapping tags between the seed paper $\mathfrak{s}$ and the sample $\mathfrak{p}$
    end for
    Calculate AUC for the classification problem $\mathbf{y} \sim \mathbf{t}$
end for
return The set of AUCs.    ▷ The average AUC is the consistency measure, while the variation provides the error estimate
```

Tagging source	Tags per paper		Coverage
	Mean	Median	
MANUAL	5.96	6	0.73
NEJI	12.22	4	0.55
DNORM	0.85	0	0.35
GNAT	0.18	0	0.08
MANUAL + NEJI	17.30	10	0.77
MANUAL + DNORM	6.74	6	0.74
MANUAL + GNAT	6.14	6	0.73
MANUAL + NEJI + DNORM + GNAT	18.26	11	0.77
NEJI + DNORM + GNAT	13.25	5	0.60

Table 1: Tags coverage by different sources. The last column shows the fraction of papers with at least one tag.

were mapped using NEJI [Campos et al., 2013]. The coverage of papers by different tagging sources is shown in Table 1.

The number of seed papers n in Algorithm 1 was chosen to be $n = 100$. Based on the preliminary experiments we chose the following hyperparameters which produced a good convergence of the measure: the number of cited papers per see paper $k = 10$, the number of negative samples per cited paper $m = 2$. The date granularity chosen was year-month: two papers were considered published at the same time if their years and months of publication in PubMed coincided.

The baseline to compare the results against was constructed by randomizing the tag sets. We linked each paper (seeds, references, and negative samples) to the tags from all the sources, and then randomly shuffled the papers, so each paper got a set of tags from some other paper in our sample. We expect the average AUC for this random tag set to be 0.5, reflecting the lack of discrimination between the cited papers and negative samples.

We calculate the AUC for each of our seed papers, and report the distribution of the results in Tukey "box and whiskers" plots.

On Figure 2 we show the consistency measure for the tagging sources: DNORM, GNAT, MANUAL, NEJI as well as the combined automatic taggers (NEJI + DNORM + GNAT).

Adding machine-generated tags to the manual ones is explored on Figure 3. Here we take manually created tags and add machine-generated ones from different sources, again using random tags as a baseline.

On Figure 4 we show cross-consistency between the manual tags and NEJI-generated ones (since GNAT and DNORM are used to predict only specific concepts like genes and diseases, they are omitted from the experiment). We used one source for tagging seed papers, and another source for tagging samples (cited papers and negative samples).

4. Discussion and conclusions

First, there is clear difference between the consistency of the randomly generated tags and the real ones (Figure 2). As expected, the consistency of the random tags is concentrated at AUC = 0.5, with some outliers both above and below this value. In contrast, the consistency of the real tags is almost always above AUC = 0.5. An exception is tagging sources of low coverage like GNAT (see Table 1), where consistency is close to 0.5. Obviously when the coverage is low, most positive and negative samples have zero overlap with their seed papers, which lowers AUC. Unexpectedly, the consistency of high coverage machine generated sources like NEJI is on par with the human tags.

Tags augmentation is explored on Figure 3. As expected, adding random tags to the manually generated ones does not noticeably change the consistency of the result. However, adding "real" machine generated tags is improving our measure, which is another evidence that the measure itself is reasonable.

The cross-consistency between manual tags and machine-generated ones is shown on Figure 4. Here we used different sources for seed papers and for samples. While cross-consistency is lower than the internal consistency of each tagger, is still is significantly higher than for random tags.

In conclusion, a simple measure of consistency of tagging: whether it is predictive for citation links in a knowledge graph,—seems to be informative about the tagging process and can be used, along with other measures, to assess and evaluate it. Cross-consistency between different taggers can be used to estimate their similarity, especially when some taggers (e.g. manual tagging) are too expensive to run on a large set of papers.

Acknowledgments

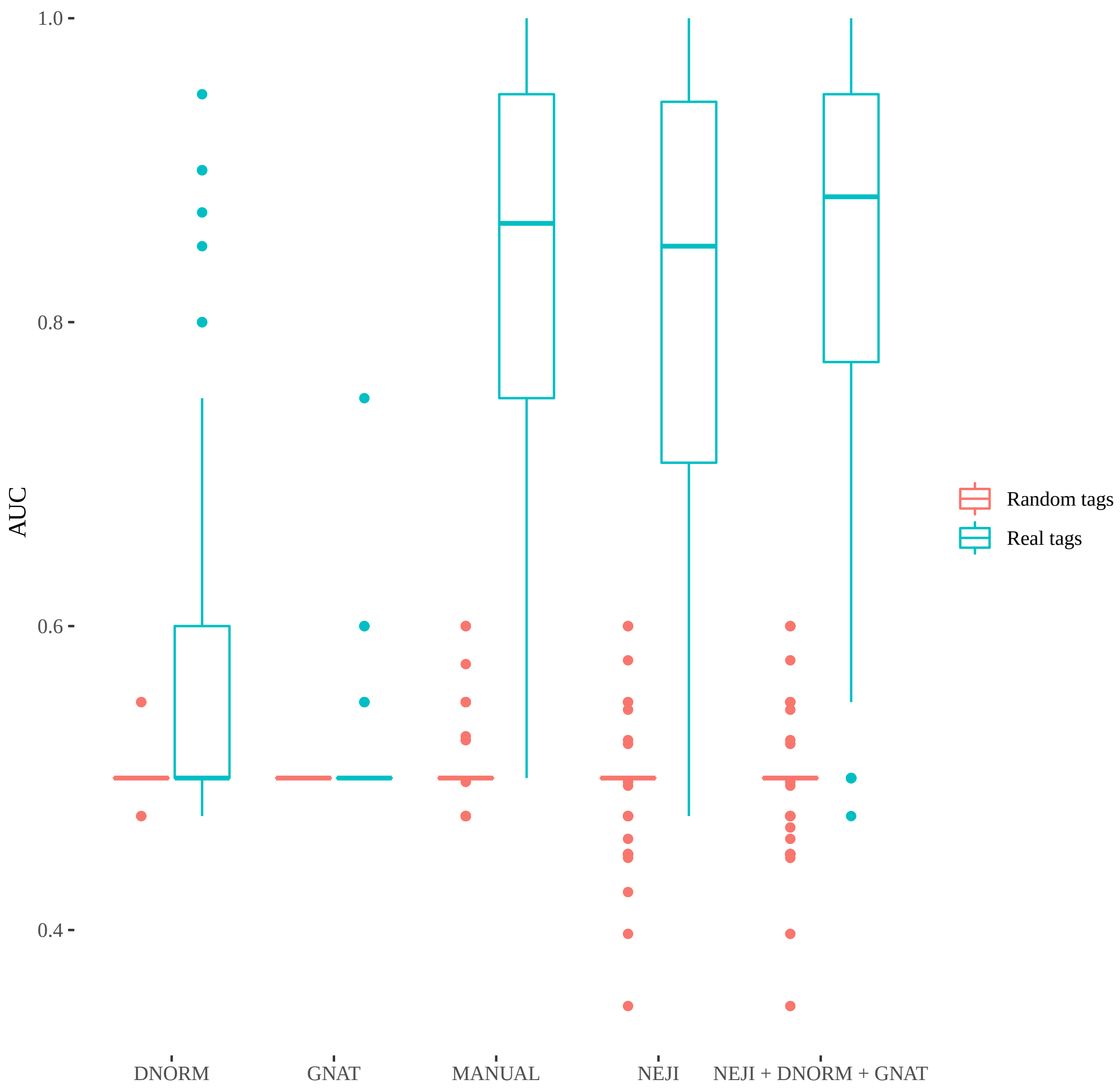

Figure 2: Consistency measure for tagging sources

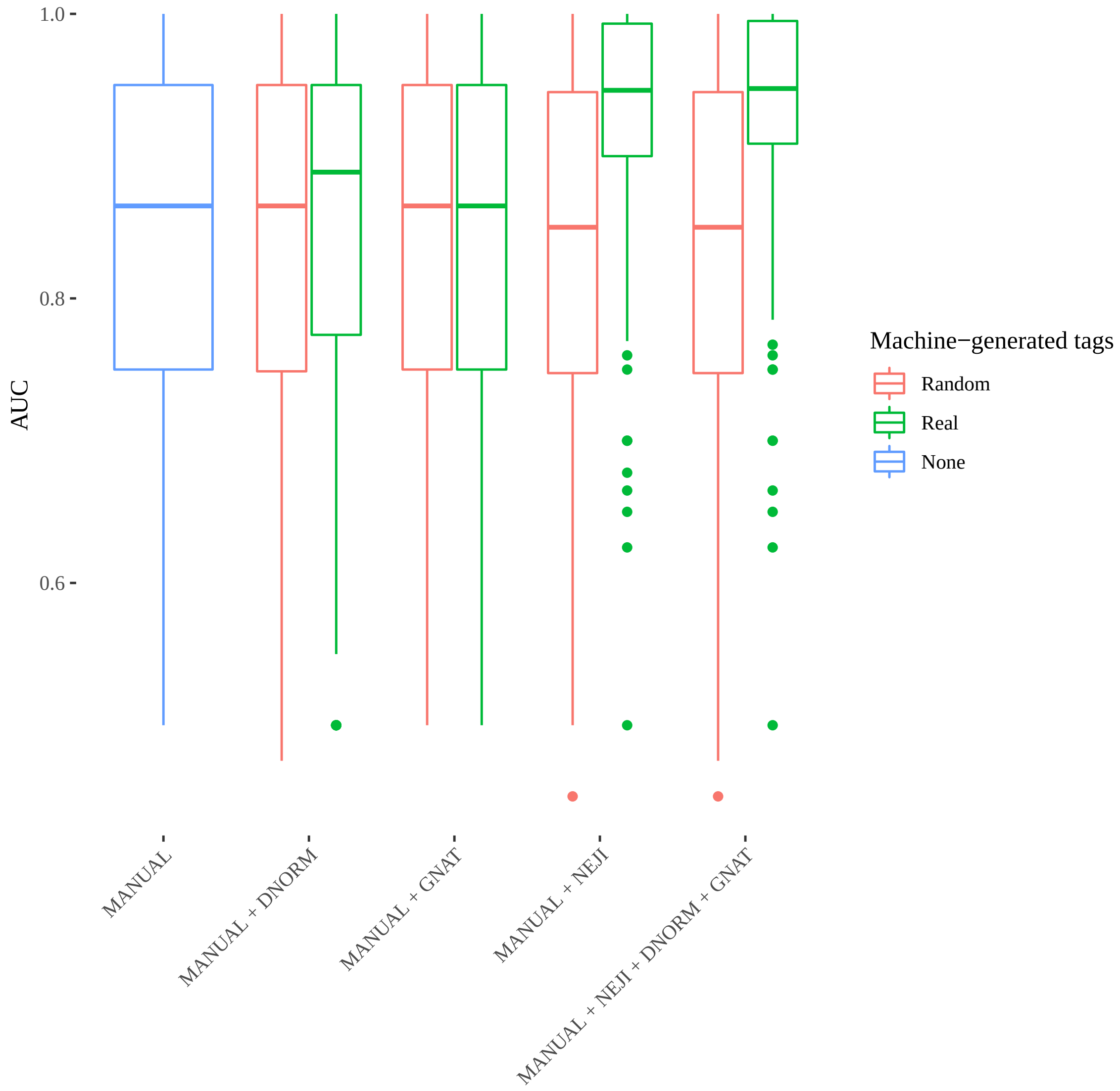

Figure 3: Consistency measure for adding machine-generated tags to manually created ones

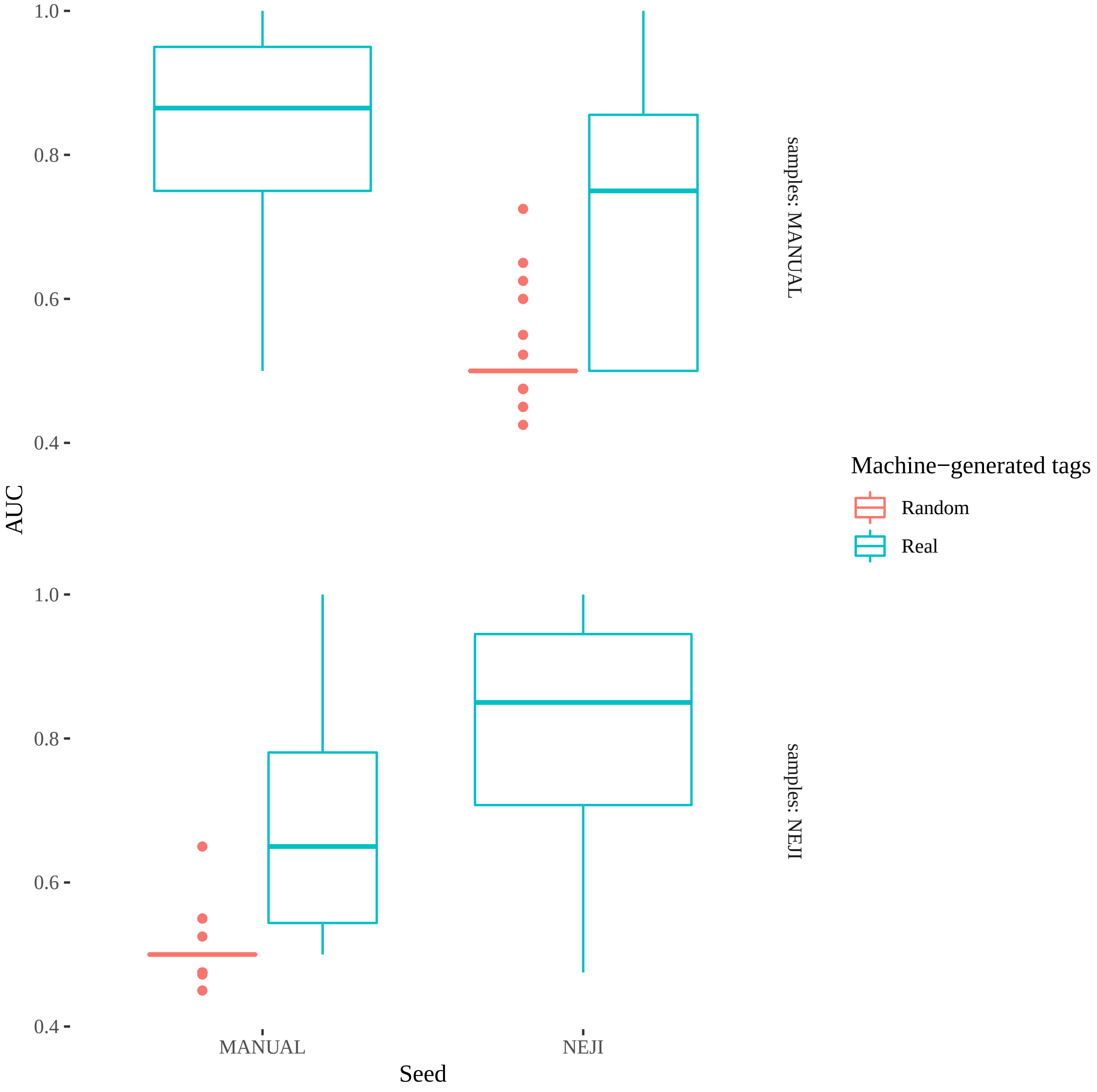

Figure 4: Cross consistency between manual tags and NEJI generated ones. X axis shows the source for the seed papers, Y axes shows the source for samples

Omitted for blind review.

References

Jorge Luis Borges. *Ficciones*. Editorial Sur, Buenos Aires, 1944.

David Campos, Sérgio Matos, and José Luís Oliveira. A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14(1):281, September 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-281. URL https://doi.org/10.1186/1471-2105-14-281.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2005.10.010. URL http://www.sciencedirect.com/science/article/pii/S016786550500303X. ROC Analysis in Pattern Recognition.

Aditya Grover and Jure Leskovec. Node2Vec: scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939754. URL http://doi.acm.org/10.1145/2939672.2939754.

Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M Bergman. The GNAT library for local and remote gene mention normalization. *Bioinformatics (Oxford, England)*, 27 (19):2769–2771, October 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr455. URL http://europepmc.org/articles/PMC3179658.

William L. Hamilton, Rex Ying, Jure Lescovec, and Rok Sosic. Representation learning on networks. WWW-18 tutorial, April 2018. URL http://snap.stanford.edu/proj/embeddings-www/.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013. doi: 10.1093/bioinformatics/btt474.

Christopher D. Manning. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19400-9.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. Hierarchical losses and new resources for fine-grained entity typing and linking. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018.