Iterative Refinement and Quality Checking of Annotation Guidelines — How to Deal Effectively with Semantically Sloppy Named Entity Types, such as Pathological Phenomena

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Erik Faessler
Jenny Traumüller, Susann Schröder, Kerstin Hornbostel

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Fürstengraben 30, 07743 Jena, Germany

udo.hahn@uni-jena.de

Abstract

We here discuss a methodology for dealing with the annotation of semantically hard to delineate, i.e., sloppy, named entity types. To illustrate sloppiness of entities, we treat an example from the medical domain, namely pathological phenomena. Based on our experience with iterative guideline refinement we propose to carefully characterize the thematic scope of the annotation by positive and negative coding lists and allow for alternative, short *vs.* long mention span annotations. Short spans account for canonical entity mentions (e.g., standardized disease names), while long spans cover descriptive text snippets which contain entity-specific elaborations (e.g., anatomical locations, observational details, etc.). Using this stratified approach, evidence for increasing annotation performance is provided by κ-based inter-annotator agreement measurements over several, iterative annotation rounds using continuously refined guidelines. The latter reflects the increasing understanding of the sloppy entity class both from the perspective of guideline writers and users (annotators). Given our data, we have gathered evidence that we can deal with sloppiness in a controlled manner and expect inter-annotator agreement values around 80% for PATHOJEN, the pathological phenomena corpus currently under development in our lab.

Keywords: Corpus annotation, annotation guidelines, named entity recognition, life sciences

1. Introduction

The assignment of linguistic meta data to natural language corpora is typically based on elaborate annotation guidelines. Such documents (rarely discussed in depth in the literature, for recent exceptions see (Wilbur et al., 2006), (Bada et al., 2010) and (Cohen et al., 2010)) contain operational definitions of the annotation task and instructions for human coders to consistently assign the desired linguistic meta data to natural language utterances. Usually, *iterative training* rounds are necessary to develop a shared understanding of the annotation task among the annotators. In this process, only marginal changes, if at all, of the underlying annotation guidelines are made. Both, the clarity of the guidelines as instruction texts and effective training of coding behavior on the basis of these guidelines are prerequisites for consensual annotation results.

Training human coders on semantic annotations is a particularly challenging task. Yet, the annotation of named entities and semantic relations connecting them differs in terms of their intrinsic task complexity. While *Persons*, *Organizations* or *Locations* from the newspaper domain, for instance, constitute named entity types for which fairly uncontroversial and stable annotation guidelines can be set up quite easily, other named entity types, such as the notoriously difficult to nail down *Gene/Protein* classes from the life sciences domain need much more efforts already for operational instructions (see, e.g., (Hatzivassiloglou et al., 2001)). Different degrees of annotation complexity are mirrored in different degrees of inter-annotator agreement (IAA), *ceteris paribus* (more than 10% when, e.g., IAAs for *Persons* and *Genes/Proteins* are compared).

In our current annotation efforts, we focus on a named entity class from the life sciences which is well-known for its difficulty—pathological phenomena, including diseases, for different species. Upon inspection of the documents related to this topic, it soon turned out that we were dealing with an entity class whose intension (what qualifies as a pathological phenomenon in the domain of discourse?) and extension (what constitute true mentions of intended pathological phenomena in raw texts?) were hard to specify. What constitutes a pathological phenomenon (and what not) is already hard to decide for professional pathologists, so much the more for human annotators. Obviously, e.g., an *appendicitis* is clearly a disease but is "a person with *high blood pressure*", certainly not being a disease statement, nevertheless describing a pathological phenomenon? Trying to deal with such borderline cases, we claim that pathological phenomena belong to a class of named entity types we here refer to as *semantically sloppy*.

In this paper, we will describe an outline of a methodology how to deal with semantically sloppy named entity types and exemplify our approach for *Pathological Phenomena*. Rather than training annotators on an *a priori* fixed notion of pathological phenomena which might be hard to justify we argue for an experience-based calibration of the annotation task by *iterative guideline revisions*. To keep control of a deeper understanding of the task and replicability of the annotation results we, in parallel, perform IAA-centered quality checks (coupling iterative guideline revision cycles with simultaneous measurements of IAA has also been proposed by (Wilbur et al., 2006)). The methodological framework emerging from this task setting is the focus of this

paper. As a proof of concept, this methodology will be applied to create a trustable corpus annotated for pathological phenomena, work which is on-going in our lab.

2. Domain Briefing: Pathological Phenomena

Pathological phenomena cover a wide range of medical observations which indicate deviations from 'normal' healthy states of an organism, typically a human being. In the medical community, the least controversial subclass of pathological phenomena are known as *diseases* with more or less clearly defined deviation criteria from the 'normal' state. Examples for diseases are *"Alzheimer's Disease"*, *"Lung Cancer"*, or *"Appendicitis"*.

Annotation of canonical diseases can be based on a large variety of established disease terminologies. The *Systematized Nomenclature for Medicine-Clinical Terms* (SNOMED CT)[1] is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. Currently, SNOMED CT contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into multiple hierarchies. The *Disease* hierarchy is available under the "Clinical Finding" root node. Another widely used disease terminology is the *National Cancer Institute thesaurus* (NCIt).[2] The NCIt provides definitions, synonyms, and other information about nearly 10,000 cancers and related diseases, 8,000 single agents and combination therapies, and a wide range of other topics related to cancer and biomedical research. The *Disease* hierarchy is available under the "Diseases, Disorders and Findings" root node. The most widely used disease terminology, which also includes a wide variety of signs, symptoms, abnormal findings, complaints, etc., is the *International Classification of Diseases* (ICD), which is part of the WHO Family of International Classifications. Version 10 of ICD contains 155,000 different codes.[3] Disease terminologies are updated regularly (consider, e.g., growth areas like viral diseases) although changes are not dramatic. Note that in medical documents the mention of a disease often comes with the mention of its anatomical locus, as well as tests and treatments, including therapies, related to it.

Beyond the terminologically clear-cut borderline of diseases the sloppy part of 'non-normality' of observations begins. Clinical notes kept in electronic health records (EHRs) or research papers (such as journal publications, etc.) contain a wide range of descriptive statements that characterize disorders and other descriptions of non-normalities (e.g., descriptions of symptoms such as *"bleeding nose"* or *"high temperature"*). Furthermore, there are mentions of observations that can be considered as an abnormal sign or finding, or merely a patient's complaint (e.g., *"facial rashes"*, *"heavy coughing"* or even *"patient felt weak"*) but by no means already indicate a concrete disease. Since the diagnostic tracing (determination or exclusion) of a disease requires to refer to critical conditions and potentially indicative single observations of a patient, all these are relevant bits of information for a comprehensive disease tracking system. Descriptive statements of this weaker type constitute the broad and hard to delineate class of what we refer to as *Pathological Phenomena* which holds *Diseases* as a proper subclass.

3. Guideline Development

In this section, we briefly outline the conduct of our guideline development activities. We started by querying the MEDLINE/PUBMED Baseline Repository[4] 2011 for all records that come with an abstract and are indexed with the MeSH[5] heading "Disease" (or a subordinate heading). The query resulted in a set of over 65,000 documents that became the basis for our annotation project.

From this document set (in the following also called the "base corpus") we drew a random sample of 200 abstracts. Screening these abstracts for mentions of pathological phenomena we compiled the first draft of our annotation guidelines. Next, we trained two expert biologists on this draft and taught them how to use our annotation tool.[6] Overall, we ended up in three iteration rounds for the definition of the annotation guidelines and training of the annotators. In each iteration we collected a random sample of 50 documents from the base corpus and let both annotators mark all mentions of pathological phenomena, requesting them to strictly adhere to the respective guideline version of that iteration round.

The main messages we learned from these three iterations can be summarized as follows:

- **Guideline Refinement after Iteration 1: Thematic Annotation Scope.** We formulated a *positive list* of desired annotation topics (standard disease names, mentions of disorders, wounds and injuries, allergies, etc.) and a complementary *negative list* of non-annotation topics (e.g., causes for diseases like genetic mutations, viruses, or bacteria). It turned out that it was particularly difficult for the annotators to distinguish between pathological phenomena on the phenotype level and non-canonical genetic conditions that cause phenotypic changes or disease patterns.

- **Guideline Refinement after Iteration 2: Annotation Mention Span.** Analyzing the annotation behaviour clearly revealed that the annotators tended to switch between short and long form span annotation for pathological phenomena. Longer text spans included, e.g., additional mentions of species specifications or statements about anatomical sites in the annotations. Both phenomena occur systematically for lots of mentions of pathological phenomena, for disease mentions in particular. In addition, the annotators still

[1] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[2] http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources

[3] http://www.cdc.gov/nchs/icd/icd10.htm

[4] http://mbr.nlm.nih.gov/

[5] http://www.nlm.nih.gov/mesh/

[6] All annotations were carried out with the *Active Learning*-compatible annotation software JANE (Tomanek et al., 2007b).

did not annotate borderline cases of pathological phenomena mentions consistently (such as *"patient felt weak"*).

As a consequence, we decided to extend the annotation to a 2-category decision task, introducing two categories to be annotated, *viz.* "short" and "long" pathological phenomenon mentions, instead of the previously used singleton category. The "long" category is used for both, extended annotation spans (see below), and the annotation of borderline cases. The annotators were instructed to annotate mentions of pathological phenomena either with the category "short" (in case of a concise mention, such as disease names), or with the category "long" (in borderline cases), or with both categories in a nested way, in case of an unclear or ambiguous text span. In the latter case, the shortest possible but still pathology-indicative part of the mention should be labeled as "short", while the longest possible, yet still only pathology-indicative text span should be labeled as "long". Consider, for example, the sentence *"She had [[clumsiness]$_{short}$ in her left extremities]$_{long}$."* While *"clumsiness"* is annotated as "short" pathological phenomenon, according to our revised guidelines, a second annotation of type "long" has to be introduced that includes the short annotation and the anatomical specification *"in her left extremities"*. Long annotations have various linguistic appearances. They occur, amongst others, as prepositional phrases (as, e.g., in *"[absence of auditory canals]$_{long}$"*), coordinations (as, e.g., in *"[markedly decreased serum IgG, IgA , and IgE levels]$_{long}$"*), and even entire sentences (as, e.g., in *"[The mass is compressing her trachea.]$_{long}$"*.

Our decision to annotate raw text based on the short – long category split is motivated by two considerations. First, there is no undisputed ground truth whatsoever concerning the 'true' textual extension of a pathological phenomenon statement. Rather we believe that the two categories meet a well-justifiable distinction. Whenever possible, annotators shall encode clear and fully evident cases of pathological phenomena mentions (e.g., *"Alzheimer's disease"*) with first-order priority. Whenever this is not possible, annotators shall encode the longest stretch of text that carries a statement clearly related to pathological phenomena (long category) with second-order priority and, in addition, mark in the long stretch any occurrences of concise pathological phenomena (short category).

While this distinction does not resolve the intrinsic problems of proper additional splits (e.g., further distinguishing the locus, symptoms, tests, etc. from the disease proper in a long span) it does, however, resolve the issue to demarcate stretches of text that are relevant for pathological phenomena and those that are definitely not. This idea should then be reflected in higher IAA values on this two-way distinction compared with previous annotation studies dealing directly with more detailed annotations.

Second, the two-way category system should allow researchers in follow-up studies to focus on their specific pathological phenomena interests since they only have to take into account the long stretch annotations. This way, our annotation exercise provides some sort of layered annotation where the long layer is open to further refinement.

In order to impose stringent quality control on the proposed procedure we measured inter-annotator agreement over several iterations of annotations, each round with different annotation guidelines. IAA was measured token-wise using Siegel & Castellan's κ (Siegel and Castellan, 1988). In the first two iterations, only one annotation category was used. In this case, κ is determined by the number of tokens for which the annotators agreed on annotating or not annotating them. After the second iteration, two annotation categories were used ("short" *vs.* "long"), and nested annotations were allowed. Consequently, for measuring IAA we had to define what we count as token-wise agreement between the annotators. We decided to consider as agreement (1) if both annotators assigned the same annotation categories (multiple assignments of the same category are not conflated) (2) if both annotators assigned at least once the category "short", and (3) if no annotation was assigned at all.

We also measured IAA ignoring the distinction between annotation categories and called the resulting value κ_{relaxed}. The κ_{relaxed} value is determined by the number of tokens that have been annotated by both annotators with at least one category, or that have not been annotated at all. The κ and κ_{relaxed}-values measured for all three iterations are summarized in Table 1.

Iteration	κ	κ_{relaxed}
1	0.71	0.71
2	0.73	0.73
3	0.80	0.85
AL	0.62	0.76

Table 1: Inter-annotator agreement of the three iterative annotation rounds and an active learning (AL) round using Siegel & Castellan's κ. The value κ_{relaxed} results from ignoring the distinction between annotation categories (short *vs.* long). (Note that in Iteration 1 and 2 we only used one annotation category, resulting in the same values for κ and κ_{relaxed}.)

As Table 1 reveals the first three iterations provide evidence for the increasing approximation of a common understanding among the annotators concerning the sloppy named entity type *Pathological Phenomena*, with IAA moving from 0.71 via 0.73 to 0.80. Values for κ_{relaxed} even peak at $\kappa = 0.85$ (Iteration 3). These numbers indicate that semantically sloppy named entities can be dealt with in an effective way given discriminative thematic and mention span criteria. After the third iteration, IAA values had reached a reasonably high level ($\kappa = 0.80$ and $\kappa_{\text{relaxed}} = 0.85$). Thus, the guidelines were frozen for further annotation.

After the empirically justified stabilization of the annotation guidelines, for the production phase of our corpus, we shifted from common random selection of utterances to be annotated to an annotation strategy based on *Active Learning* (AL). AL recently has created a lot of interest in the annotation community because of its promise to speed up

the annotation process at (almost) no negative costs in terms of annotation quality (Tomanek et al., 2007a). AL targets the most uncertain and thus most tricky annotation samples and selects it for annotation by the annotators. We applied this biased sampling technique to the base corpus (composed of 530,000 sentences). For concrete annotation work we used the JANE system, a dedicated annotation software framework developed at our lab (Tomanek et al., 2007b).

In order to screen the quality of annotations using the AL approach, we also measured IAA on a document set from five subsequent AL rounds, where each round contained 50 sentences to be annotated (thus, the overall set contains 250 sentences). The IAA results reveal that agreement values drop for the AL approach—0.62 considering both categories and 0.76, if we consider annotation spans only (see Table 1). This result comes at no surprise because an AL system, in general, always prefers samples with a much higher annotation complexity than encountered in a random-based selection approach. Thus, agreement should be predictably lower than in the random-selection case which selects complex but also 'easy' annotation instances.

4. Discussion

In terms of a shaping methodology for dealing with semantically sloppy named entity types, such as *Pathological Phenomena*, we propose two main steps. First, clear thematic demarcation lines have to be drawn to single out annotation-relevant from annotation-irrelevant language data by supplying thematically positive and negative lists, respectively. Second, strict and lazy mention span criteria have to be defined in case of span-ambiguous textual mentions of semantically sloppy named entities, while in non-ambiguous cases the standard annotation procedure for named entities still applies.

These principles have emerged from observations of the linguistic appearance of pathological phenomena. We found that standard diseases, a proper subset of pathological phenomena, are *named*, while pathological phenomena tend to be verbally *described* on various axes, e.g., involving species characteristics or anatomical regions and locations. Since we intend to leave it to the development of future named entity taggers and relation extractors to differentiate between such properties (species, locations), we have opted *a priori* for a perspective on annotation which focuses on the textual extension of descriptions rather than pre-structure a set of complex entities and relations.

As an example of the latter type of annotation strategy consider the CLEF (Clinical E-Science Framework) corpus (Roberts et al., 2009) which is composed of clinical narratives, histopathology reports, and imaging reports from 20,000 cancer patients. For each of these three genres, 50 documents were meticulously annotated with several disease-specific types of clinical entities, namely *Condition* (including symptom, diagnosis, complication, conditions, problems, functions, processes, and injury), *Result* (the numeric or qualitative finding of an investigation, excluding *Condition*), and *Locus* (the anatomical structure or location, body substance, or physiological function, typically the locus of a *Condition*). Very often, *Conditions* are mentioned in relation to *Locus* as, for example, in "*[melanoma]*$_{Condition}$ *located in [groin]*$_{Locus}$" or "*[left breast]*$_{Locus}$ *[cancer]*$_{Condition}$." Furthermore, several relation types are annotated, including *HasFinding*, *HasIndication*, *HasLocation*, *HasTarget*, and *Modifies*, as well as temporal annotations (such as *Before*, *After*, *Overlap*, *Includes*) for time-sensitive named entities. Thus, the annotation process for diseases is broken down into the annotation of many fundamental clinical and anatomical entities and their relationships. A wide range of IAA scores are reported for such a relational decomposition of annotation (ranging, e.g., from 29% to 95% at the named entity level for different types types of clinical documents) which suggests that this fine-grained relationship annotation for clinical entities is a really hard task (Roberts et al., 2007). The latest round of the i2b2 Challenge (Uzuner et al., 2011) led to the creation of an entity/relationship corpus also made of clinical documents, which is similar in thematic scope though not comparable in annotation depth to the CLEF effort.

A far more restricted perspective on the pathological phenomenon annotation task underlies the Disease Corpus from EBI (Jimeno et al., 2008) or the Arizona Disease Corpus (AZDC) (Leaman et al., 2009). Both only deal with *Disease* type annotations, a proper subset of *Pathological Phenomena*. The EBI corpus contains 600 sentences from the Online Mendelian Inheritance in Man (OMIM) database,[7] for which an IAA of 0.51 kappa (which is low, even by biomedical standards) is reported for two annotators. AZDC provides 3,228 disease mention annotations (1,202 unique disease names) for 2,856 MEDLINE abstracts. Mentions of organisms and species are explicitly excluded from the disease annotation span. So for "*human insulin-dependent diabetes mellitus*", the span "*insulin-dependent diabetes mellitus*" would be annotated as *Disease*. Furthermore, there exist several corpora which deal with particular disease types, such as the PENNBIOIE ONCOLOGY corpus,[8] which is composed of 1,414 MEDLINE abstracts annotated for the molecular genetics of cancer.

In a similarly over-constrained way, disease-focused corpora have been annotated using established terminologies as the target tag language. (Ogren et al., 2008) report on a corpus which contains 1,556 annotations on 160 clinical notes using 658 unique concept codes from SNOMED-CT corresponding to human disorders. IAA for four annotators is reported, among others, for span (0.91) and mapping to concept code (0.82). In earlier work, (Pestian et al., 2007) describe a clinical notes corpus composed of almost 2,000 documents annotated at the document level for billing codes (45 categories taken from the disease classification ICD-9CM).[9]

Accordingly, our approach (unlike the EBI and AZDC corpora) explicitly deals with a wide range of pathological phenomena and, thus, includes diseases as a proper subset. On the other hand, it also explicitly avoids (unlike CLEF) getting too far into fine-grained decomposition problems of named entity and relational annotations (involving locational and result-type relations). Still, our long-form

[7] http://www.ncbi.nlm.nih.gov/omim

[8] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T21

[9] http://www.cdc.gov/nchs/icd9.htm

annotation policy marks text spans which can be further investigated by suitable *Locus* and *Result* taggers, if required. On-going work in our lab deals with the creation of a corpus of pathological phenomena, called PATHOJEN. It is based on the guidelines of the third iteration and will cover about 100,000 tokens (corresponding to approximately 400 MEDLINE documents). Building on previous experience with Active Learning-based annotation, PATHOJEN will be one of the rare non-toy biomedical corpora which are systematically generated using AL-type biased sampling in order to optimise for cost-efficient, yet high quality annotation results. Once this corpus is at hand, taggers for *Pathological Phenomena* and *Diseases* will be trained along the lines of GENO (Wermter et al., 2009), a *Gene/Protein* tagger which performs on a par with the top-ranked gene/protein taggers on the BIOCREATIVE II benchmark.

Acknowledgements. This work is partially funded by a grant from the German Ministry of Education and Research (BMBF) for the *Jena Centre of Systems Biology of Ageing* (JENAGE) (grant no. 0315581D).

5. References

Michael Bada, Lawrence E. Hunter, Miriam Eckert, and Martha Palmer. 2010. An overview of the CRAFT concept annotation guidelines. In *LAW IV – Proceedings of the 4th Linguistic Annotation Workshop at ACL 2010*, pages 207–211. Uppsala, Sweden, 15-16 July 2010.

K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner, Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence E. Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of the 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM) at the 7th Language Resources and Evaluation Conference (LREC 2010)*, pages 37–41. Valletta, Malta, March 18, 2010.

Vasileios Hatzivassiloglou, Pablo A. Duboué, and Andrey Rzhetsky. 2001. Disambiguating proteins, genes, and RNA in text: A machine learning approach. *Bioinformatics*, 17(Suppl 1):S97–S106.

Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl. 3):S3.

Robert Leaman, Graciela Gonzalez, and Christopher Miller. 2009. Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark. In *LBM 2009 – Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*, pages 82–89. Seogwipo-si, Jeju Island, South Korea, November 8-10, 2009.

Philip V. Ogren, Guergana K. Savova, and Christopher G. Chute. 2008. Constructing evaluation corpora for automated clinical named entity recognition. In *LREC 2008 – Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 3143–50. Marrakech, Morocco, 29-30 May 2008.

John P. Pestian, Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, and K. Bretonnel Cohen. 2007. A shared task involving multi-label classification of clinical free text. In *BioNLP 2007 – Proceedings of the ACL 2007 Workshop on Biological, Translational, and Clinical Language Processing*, pages 97–104. Prague, Czech Republic, June 29, 2007.

Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay (Subbarao) Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. 2007. The CLEF corpus: Semantic annotation of clinical text. In *AMIA 2007 – Proceedings of the Annual 2007 Symposium of the American Medical Informatics Association*, pages 625–629. Chicago, Illinois, USA, 10-14 November 2007.

Angus Roberts, Robert J. Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Sidney Siegel and John N. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 2nd edition.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007a. An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In *EMNLP-CoNLL 2007 – Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 486–95. Prague, Czech Republic, June 28-30, 2007.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007b. Efficient annotation with the Jena ANnotation Environment (JANE). In *The LAW at ACL 2007 – Proceedings of the Linguistic Annotation Workshop*, pages 9–16. Prague, Czech Republic, June 28-29, 2007.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–6.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance gene name normalization with GeNo. *Bioinformatics*, 25(6):815–821.

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: Definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356.