The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation In The Biomedical Domain

Kevin Bretonnel Cohen, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, Lawrence Hunter

The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation In The Biomedical Domain

K. Bretonnel Cohen, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, Lawrence E. Hunter

1 Motivation For The Work

A major question in linguistics is whether theoretical accounts of the general language work for specific domains. Similarly, in natural language processing, it is clear that general-domain solutions often fail when applied to specialized domains. One such specialized domain, which is increasingly seen as crucial to understanding human biology and disease, is the biomedical domain. For this reason, biomedical corpus construction has been an area of considerable activity in recent years—for example, just in the past five years: (ordered by year of publication and then by first author's last name), [34, 63, 79, 93, 1, 33, 40, 59, 76, 72, 54, 94, 22, 32, 64, 75, 77, 96, 3, 4, 10, 17, 19, 46, 52, 67, 68, 74, 99, 7, 20, 23, 21, 28, 69, 16, 58, 57, 80, 82, 81, 95, 83, 89].

Historically, the great majority of work in biomedical natural language processing has been done using abstracts of journal articles. In contrast, the Colorado Richly Annotated Full Text (CRAFT) corpus consists entirely of full-text journal articles. The primary motivation for the annotation project was the accumulating body of evidence indicating that the bodies of journal articles contain much information that is not present in the abstracts, and that the textual and structural characteristics of article bodies are different from those of abstracts [8, 26, 90, 84, 18, 2, 48, 51, 13]. When we began the project, there was no large resource of full-text journal articles for system building or evaluation; other than the CRAFT corpus, this continues to be the case. Earlier projects on full-text biomedical journal articles are typically not manually annotated, and none of them that we are aware of have annotation of linguistic structure.

Kevin Bretonnel Cohen
Computational Bioscience Program, University of Colorado School of Medicine (Cohen, Funk, Bada, Hunter), Department of Linguistics, University of Colorado at Boulder (Cohen, Palmer), University of Melbourne (Verspoor), and University of Paris-Sorbonne (Fort) e-mail: kevin.cohen@gmail.com

For these reasons, we sought and received funding to annotate a corpus of complete journal articles. The motivation for the annotation schema, particularly the named entity annotation schema, was that although there is a large number of broad semantic classes of named entities that are of interest to biomedical natural language processing consumers, most work on named entity recognition in the biomedical domain had focussed on genes and gene products only. We hoped to enable research on other broad semantic classes of named entities by increasing the scope of the annotation project considerably, compared to previous work.

Returning to the subject of funding, the process of obtaining it was somewhat circuitous and ultimately somewhat surprising. We initially submitted a proposal related to natural language processing applied to full-text journal articles; it contained a large annotation component (about half of the budget), since there was no existing data set that our work could be evaluated on. The National Institutes of Health funded the project under the R01 funding mechanism, but declined to fund the portion of the budget that would have paid for the annotation work, on the grounds that data preparation was not research (NIH's view, not the authors'—see, for example, [9, 86, 100, 37, 29, 24, 38, 66, 47]). We were encouraged to apply for a resource development grant, and that was funded, for about the same amount of the budget as had been refused on the original R01 application. This was an unexpectedly happy outcome, but unfortunately, the National Institutes of Health no longer offer the resource development funding mechanism, and it seems unlikely that other annotation groups will be funded for similar projects. The situation might be different for clinical data.

2 Annotation Scheme

The development of CRAFT was characterized by what [88] has described as a "multi-model" annotation task. [24] characterizes these as separate Elementary Annotation Tasks (EAT). In a multi-model task, there are separate models for highly disparate elements of the task. In the typical case, there is a linguistic annotation task and corresponding model, and what [88] has characterized as a "light" annotation task, in which domain experts carry out annotation that requires domain expertise but does not require any knowledge of linguistics[1]. In the case of CRAFT, the two models were linguistic and named-entity-related—neither was "light."

The named entity annotation of the CRAFT corpus had the goal of annotating textual references to all, and to only, terms from a realist ontology [27, 85]. Seven different ontologies were used, containing more than 100,000 concepts. The task has some commonalities with the ACE [53] corpora—both annotation efforts begin with an external model of the world. It differs in that the ACE annotation uses on the order

[1] When we mention *linguistic* annotation, we mean part of speech, syntactic, structural (e.g. sentence boundaries and tokenization) and coreference annotation. This is contrasted with *named entity* annotation, referred to more broadly as 'semantic' annotation when we refer to broad semantic categories, such as Sequence Ontology concepts or NCBI Taxonomy entities.

of tens of semantic classes of entities. Like WordNet—another large, hierarchically-structured vocabulary (see, for example, [39] for an in-depth discussion of compositionality in WordNet)—realist ontologies may contain many concepts that cannot be expressed in a single word (e.g. GO:0032332 *positive regulation of chondrocyte differentiation*), as well as many terms that contain other terms (for example, *chondrocyte differentiation* is also a term in the ontology), making recognizing those concepts in text somewhat different from recognizing a word in English text and more like recognizing MUC-style named entities [30, 11] due to the boundary and overlapping mentions issues [36, 24]. The design of the named entity annotation schema and process was broadly similar to other annotation projects. Contrary to the approach used in the linguistic annotation, there was no facility for an annotator to formally mark instances about which they had questions or that needed to be returned to; these were instead handled by a formal weekly discussion process.

The potential uses of the annotation project were broadly construed to be applications in natural language processing and in theoretical linguistics. These potential uses of the annotations did not particularly influence the development of the structural annotation guidelines, which were mostly adapted from other projects. However, specific considerations of biomedical use cases did influence the development of the named entity annotation model and guidelines quite a bit. In particular, in the biomedical community, there is an enormous need to not just be able to recognize strings in text that represent some broad semantic class, but to be able to map those strings in text to specific entries in a database or concepts in an ontology or controlled vocabulary. Thus, our annotation model and guidelines were heavily focussed on this "normalization" issue [55, 56, 49].

The selection of annotation guidelines for the coreference annotation was overtly political, in that a deliberate choice was made to align with the guidelines of some other project, rather than creating new guidelines. After considering a number of sets of guidelines, the OntoNotes guidelines created by BBN [71, 70] were adopted, with minor changes and additions that did not affect compatibility with the OntoNotes data. [14] describes the reasoning behind the choice of the OntoNotes guidelines.

Development of the annotation schema affected the development of linguistic knowledge only in very small ways, specifically with respect to the types of morphosyntactic entities that were represented. Minor additions to the Penn Treebank guidelines [50] had to be made in order to account for predicators that are represented in biomedical text but not in the materials of the Penn Treebank. They are described in [98]. The model for the linguistic annotation was not substantially different from typical treebanking efforts and will not be described in much further detail, beyond noting that a small number of additional phrasal categories needed to be added, as well as some changes to our conception of how to represent formulae (see [98] for details).

The model for the named entity annotation was as follows. Following [73], we consider an annotation model as a triple $M = T,R,I$, where

- M = **M**odel
- T = Vocabulary of **t**erms
- R = **R**elation between terms

- I = **I**nterpretation of terms

Then,

T = {Concept, Gene_Ontology_concept, Cell_Type_Ontology_concept, ChEBI_concept, NCBI_Taxonomy_concept, Protein_Ontology_concept, Sequence_Ontology_concept, Entrez_Gene_entry}

R = {Concept ::= Gene_Ontology_concept, Cell_Type_Ontology_concept, ChEBI_concept, NCBI_Taxonomy_concept, Entrez_Gene_entry}

I = {Gene_Ontology_concept = "list of all concepts in the Gene Ontology;" similarly, for the other ontologies and vocabularies.}

2.1 Materials

2.1.1 Sampling

The sampling method was based on the goal of ensuring biological relevance. In particular, the sample population was all journal articles that had been used by the Mouse Genome Informatics group as evidence for at least one Gene Ontology or Mouse Phenotype Ontology "annotation," in the sense in which that term is used in the model organism database community. In the model organism database community, it refers to the process of mapping genes or gene products to concepts in an ontology, e.g. of biological processes or molecular functions—see [12] for the interacting roles of model organism database curation and natural language processing.

2.1.2 Inclusion criteria

The inclusion criteria were that an article had to have been used as evidence for at least one Gene Ontology annotation, had to be available with an open access license (which is crucial to being able to distribute the data [25]), and had to be available in the PubMed Central XML format (which is crucial to it being amenable to annotation). 97 documents in the sample population met these criteria.

2.1.3 Exclusion criteria

There were no exclusion criteria, other than failure to meet the inclusion criteria. All documents that met the inclusion criteria were included in the corpus.

2.1.4 Balance and representativeness

The resulting document collection is probably not balanced, as there was not a large enough set of documents meeting the inclusion criteria to apply any principled ap-

proach to the selection of contents. On the other hand, it probably *is* representative of the domain, in that a broad variety of topics within the very broad field of mouse genomics are represented—development, physiology, genetics, disease, etc. The representativeness of CRAFT is further supported by the low Kullback-Leibler divergence between CRAFT and other biomedical corpora, as calculated from lexical distributions [97]. The lexical distributions generally follow the patterns that would be expected in a sublanguage corpus [92].

3 Physical Representation

The annotated data was generated with a variety of tools, some of which were used for the linguistic annotation and some of which were used for the named entity annotation.

The linguistic annotation is represented in the widely known Penn Treebank format [50], with the addition of a small number of tags and phrasal categories to accommodate the idiosyncrasies of the domain (see above). This representation was chosen due to its wide familiarity to the corpus linguistics and natural language processing communities.

The primary representation for the named entity annotation is the Knowtator format [61, 62]. This representation was chosen because the Knowtator annotation tool is optimized for use in annotation with ontologies as elements of the annotation model, and the annotation effort involved ontologies of the biomedical domain quite heavily.

Disadvantages of this representation are that it is unfamiliar to the community and difficult to manipulate computationally. For that reason, the primary representation was converted to a number of other formats, including GENIA-style XML [43, 65, 44, 91] and brat [87]. (One early adopter of the corpus did not like any of these representations and converted the annotations to a set of tab-separated values.) More recently, CRAFT has been integrated into the PubAnnotation project [45, 42, 41] and converted to the JSON-LD format (Sampo Pyysalo, personal communication).

4 Annotation Process

The annotation process was quite different for the linguistic annotation, named entity annotation, and coreference annotation.

The linguistic annotation was done by linguists–typically graduate students. It was carried out using conventional, broadly accepted methodologies, such as were used in the creation of the Penn Treebank [50]. Annotators were trained until they could achieve about 80% inter-annotator agreement on previously annotated materials. (Inter-annotator agreement was calculated as F-measure, using the precision

and recall values from the `evalb` bracket scoring program with a modified version of the Collins parameter file.) They then participated in double-blind annotation by multiple annotators, with resolution of disagreements by a senior annotator. Materials were pre-tagged with lexical categories (using the GENIA parser) and syntactic structure (using the OpenNLP parser), and these automatic annotations were reviewed and corrected by the human annotators.

The named entity annotation was done by PhD students and PhDs in the biological sciences. It was carried out in a single-blind fashion, with checking by a senior annotator. Inter-annotator agreement numbers that are reported were calculated as F1 between the blind annotator and the senior annotator's corrections. There was limited automatic pre-annotation, done by string-matching for some of the ontologies.

The coreference annotation was done by a combination of linguistics graduate students and biological science graduate students, with resolution of disagreements by a linguist. We did not note any obvious differences in performance between the linguists and biologists, although we did not look for such differences closely. Because the coreference annotation was being done at the same time as the syntactic annotations, annotators did not have access to gold-standard syntactic structures to use in the annotation process. In retrospect, we should probably have used an automatic chunker before the coreference annotation, as this probably would have increased our inter-annotator agreement, even if imperfect [35].

5 Evaluation/Quality Control

The main mechanisms for quality control in CRAFT were monitoring inter-annotator agreement [5], and in the case of the linguistic annotations, double-blind annotation with resolution of inter-annotator disagreements.

In the case of the named entity annotations, inter-annotator agreement was measured approximately weekly. As described above, the majority of the named entity annotation was single-blind annotation with correction by the lead annotator, so inter-annotator agreement is actually correctness of the initial annotator as judged by the lead annotator. The inter-annotator agreement statistics are broken down by broad semantic category in Figures 1 and 2. As can be seen, the inter-annotator agreement fluctuates wildly, but converges toward a high value fairly quickly. This is consistent with the annotator learning curve described in [50] for the Penn Treebank.

As an additional quality control check for the syntactic annotations, the CorpusSearch tool was used to validate the tree banking. About 150 CorpusSearch queries were written to search for a variety of common error types, such as phrasal/part of speech mismatches (e.g. a phrase marked as a prepositional phrase that does not actually have a preposition).

Despite the syntactic quality control efforts, some bugs seem to have snuck through in the format, since we have not been able to run the `tprep` program suc-

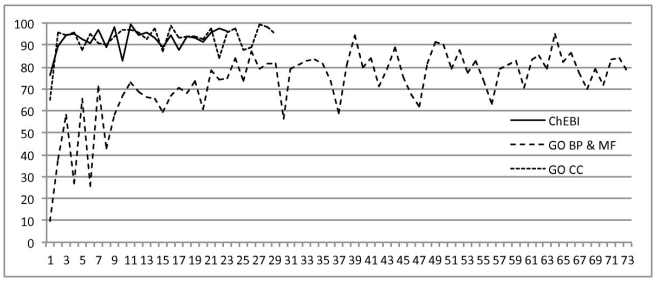

Fig. 1 Change in inter-annotator agreement over time for the ChEBI, Gene Ontology biological process and molecular function, and Gene Ontology cellular component ontologies. The *y* axis is inter-annotator agreement and the *x* axis is cumulative weeks of effort on the project. Figure from [6].

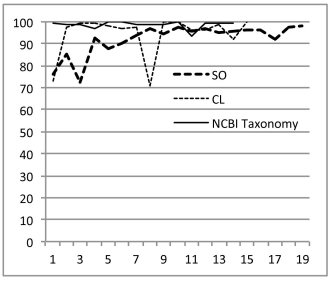

Fig. 2 Change in inter-annotator agreement over time for the Sequence Ontology, Cell Line Ontology, and NCBI Taxonomy. The *y* axis is inter-annotator agreement and the *x* axis is cumulative weeks of effort on the project. Figure from [6].

	Annotator-Annotator agreement			Annotator-Gold agreement		
	A1-A2	**A1-A3**	**A2-A3**	**A1**	**A2**	**A3**
Precision	90.58	90.18	90.13	94.98	94.58	94.39
Recall	91.02	92.31	89.39	95.92	94.98	93.16

Table 1 Inter-annotator agreement and annotator-gold standard agreement for the syntactic annotation. Adapted from [98].

cessfully on the data, rendering it inaccessible to syntactic search tools like `tgrep`. We had similar problems trying to map the syntactic annotations to JSON, and were able to isolate them to specific files. The syntactic annotation for the majority of the files is searchable through the PubAnnotation SPARQL interface.

Finally, quality was assessed by attempting to train machine learning models on the corpus. It was found that high-performing models could be trained on the linguistic annotations, although much of the named entity annotation was too sparse to allow for training a good model [98]. This is consistent with high quality for the linguistic annotations [73].

6 Some characteristics of the corpus and of the task

An initial assumption in the design of the named entity annotation was that there would be serious issues related to the length of terms in the ontologies (as measured in words). A post hoc analysis of the ontologies and the annotations showed that while there is some variability both in the lengths of the terms of the concepts from the ontologies that the annotators actually annotated and the corresponding text spans in the corpus (see Figure 3), on the whole both the terms in the ontologies and the corresponding text spans were relatively short. The terms associated with the concepts in the ontologies that we actually annotated had a mean length of 2.4 words and median length of 2.0 words, with a standard deviation of 1.34 words. The corresponding text spans that the annotators selected had a mean length of 2.0 words, median length of 1.0 word, and a standard deviation of 1.52 words. The means were statistically significantly different by two-tailed t-test, p-value $<$ 2.2e-16.

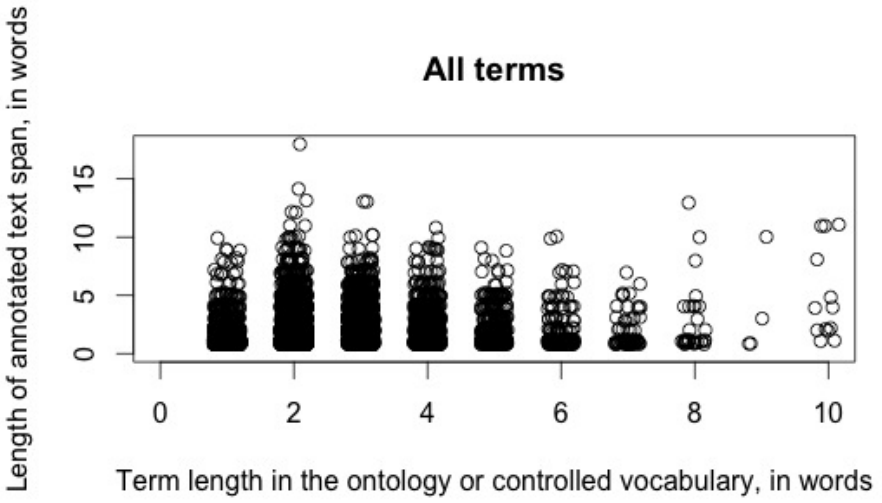

Fig. 3 Length of annotated text span, in words, over corresponding term length in the ontologies and controlled vocabularies, in words. A data point at 10 on the *x* axis and 1 on the *y* axis means that a 10-word term in some ontology corresponded to a 1-word span in the corpus. All ontologies and vocabularies are combined in this figure. The R `jitter()` function is used to reduce data points overlaying each other.

[24] proposes a model for evaluating the complexity of manual annotation tasks. It considers a project in relation to the levels of discrimination (identification) of the elements to annotate, boundary delimitation, expressiveness of the annotation language, tagset dimension, degree of ambiguity, and context. The CRAFT named

entity annotation task is an interesting case study for the model. Comparing the CRAFT task to another named entity annotation task, the Quæro structured named entities annotation task [31], CRAFT varies quite a bit on several complexity dimensions, even without performing the full calculation of the complexity dimensions.

The most obvious dimension of contrast is the tagset dimension measure. For CRAFT, it clearly gets a score of 1.0 (on a scale of 0.0 to 1.0), whereas the Quæro task has a score of only 0.34, due to the much smaller tagset. Also, the context that must be taken into account was much larger in CRAFT, as the annotators sometimes needed to read the whole text to be able to perform the task (1.0 complexity as compared to 0.75 in Quæro). Both projects used a type language, so the expressiveness is the same (0.25). The ambiguity of the CRAFT project cannot be evaluated without more complex calculations, as there were no traces left by the annotators when they had questions. Finally, as some entities were pre-annotated, the discrimination and delimitation dimensions should be somewhat less complex, as the Quæro corpus was not pre-annotated.

Overall, this rough analysis using the model described in [24], comparing the CRAFT project and a similar project, helps to elucidate the task in terms of factors that contribute to its complexity: a hugely complex tagset, a large context to take into account, and the utility of explicit ambiguity traces. Using this model beforehand, at the onset of the annotation campaign, could have helped to highlight those issues, and design the task a bit differently. For example, an analysis of the complexity of the task suggests that it might be helpful to simplify the tagset or to allow annotators to use a parent node.

7 Usage

The annotated corpus is available under a very permissive Creative Commons Attribution 3.0 (CC BY) license, on the SourceForge web site. It is freely available to any user. The initial release comprised 70% of the data. The rest has been held out for use in shared tasks and will be released in two increments of 15%.

So far, the data has been used for named entity recognition projects. The Cocoa system [78] appears to have been evaluated against the CRAFT Entrez Gene, Protein Ontology, and Sequence Ontology annotations. The BeCAS system [60] was evaluated on all of the broad semantic classes in CRAFT. It is not known how much contribution the linguistic annotation makes to machine learning for these named entity recognition tasks, as no ablation experiments have addressed this question thus far. The data has also been incorporated into the PubAnnotation project [45, 42, 41].

8 Discussion and conclusions

Our experience with building the CRAFT corpus suggests that multi-model annotation task definitions can scale to large projects. The heuristic of "always giving the annotators a way out" (Martha Palmer, personal correspondence) was valuable in the linguistic annotation work.

Four years after the publication of the first paper on CRAFT, the reference ontologies that constituted the interpretation in the named entity annotation model have changed, as they constantly do. This is not fatal to the utility of the corpus, as the versions of the ontologies that we used are easily available through their archiving systems. The basic structure of the concept normalization task that the annotations were meant to support does not change, nor does the basic structure of the ontologies. However, our experience with another resource that we prepared for evaluating concept normalization systems [15] suggests that users will want to see updated annotations, and we are actively engaged in that task. It remains to be seen if continuing to do this without explicit funding for maintenance is a sustainable model.

Acknowledgements The authors gratefully acknowledge the contributions to this work of the annotators, especially lead annotator Arrick Lanfranchi; Colin Warner for help with reconstructing the quality assurance approach; Amber Stubbs for discussion of multi-model and light annotation tasks; Paul Foster for help with Devanagari; and BBN for use of their coreference annotation guidelines.

References

1. Abacha, A.B., Zweigenbaum, P.: Annotation et interrogation sémantiques de textes médicaux. Atelier Web Sémantique Médical, IC (2010)
2. Agarwal, S., Yu, H.: Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinformatics **25**(23), 3174–3180 (2009)
3. Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W.F., Warner, C., Hwang, J.D., Choi, J.D., Dligach, D., Nielsen, R.D., Martin, J., et al.: Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association (2013)
4. Ambert, K.H., Cohen, A.M., Burns, G.A., Boudreau, E., Sonmez, K.: Virk: an active learning-based system for bootstrapping knowledge base development in the neurosciences. Frontiers in Neuroinformatics **7** (2013)
5. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
6. Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Jr., W.A.B., Cohen, K.B., Verspoor, K., Blake, J.A., Hunter, L.E.: Concept annotation in the CRAFT corpus. BMC Bioinformatics **13**(161) (2012)
7. Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P.C., Erickson, B., Miller, T., Lin, C., Savova, G., Pustejovsky, J.: Temporal annotation in the clinical domain. In: Proceedings of the Association for Computational Linguistics, pp. 143–154 (2014)

8. Blaschke, C., Valencia, A.: Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. Comparative and Functional Genomics **2**(4), 196–206 (2001)
9. Boguraev, B., Ide, N., Meyers, A., Nariyama, S., Stede, M., Wiebe, J., Wilcock, G. (eds.): Proceedings of the Linguistic Annotation Workshop. Association for Computational Linguistics, Prague, Czech Republic (2007). URL http://www.aclweb.org/anthology/W/W07/W07-15
10. Castro, L.G., McLaughlin, C., Garcia, A.: Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. Journal of Biomedical Semantics **4**(Suppl 1), S5 (2013)
11. Chinchor, N., Robinson, P.: Muc-7 named entity task definition. In: Proceedings of the 7th Conference on Message Understanding, p. 29 (1997)
12. Cohen, K.B.: BioNLP: Biomedical text mining. In: N. Indurkhya, F.J. Damerau (eds.) Handbook of natural language processing, 2nd edition (2010)
13. Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C., Hunter, L.E.: The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinformatics **11**(492) (2010)
14. Cohen, K.B., Lanfranchi, A., Corvey, W., Jr., W.A.B., Roeder, C., Ogren, P.V., Palmer, M., Hunter, L.E.: Annotation of all coreference in biomedical text: Guideline selection and adaptation. In: BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining, pp. 37–41 (2010)
15. Cohen, K.B., Roeder, C., Jr., W.A.B., Hunter, L., Verspoor, K.: Test suite design for biomedical ontology concept recognition systems. In: Proceedings of the Language Resources and Evaluation Conference (2010)
16. Collier, N., Paster, F., Campus, H., Tran, A.M.V.: The impact of near domain transfer on biomedical named entity recognition. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL, pp. 11–20 (2014)
17. Collier, N., Tran, M.V., Le, H.q., Ha, Q.T., Oellrich, A., Rebholz-Schuhmann, D.: Learning to recognize phenotype candidates in the auto-immune literature using svm re-ranking. PloS ONE **8**(10), e72,965 (2013)
18. Corney, D.P., Buxton, B.F., Langdon, W.B., Jones, D.T.: BioRAT: extracting biological information from full-length papers. Bioinformatics **20**(17), 3206–3213 (2004)
19. Dai, H.J., Wu, J.C.Y., Tsai, R.T.H.: Collective instance-level gene normalization on the IGN corpus. PLoS ONE **8**(11), e79,517 (2013)
20. Doğan, R.I., Comeau, D.C., Yeganova, L., Wilbur, W.J.: Finding abbreviations in biomedical literature: three BioC-compatible modules and four BioC-formatted corpora. Database **2014**, bau044 (2014)
21. Doğan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics **47**, 1–10 (2014)
22. Doğan, R.I., Lu, Z.: An improved corpus of disease mentions in PubMed citations. In: Proceedings of the 2012 workshop on biomedical natural language processing, pp. 91–99. Association for Computational Linguistics (2012)
23. Doğan, R.I., Wilbur, W.J., Comeau, D.C.: BioC and simplified use of the PMC Open Access dataset for biomedical text mining. In: Proceedings of the 2014 Workshop on Biomedical Text Mining, Language Resources And Evaluation Conference (2014)
24. Fort, K., Nazarenko, A., Rosset, S.: Modeling the complexity of manual annotation tasks: a grid of analysis. In: Proceedings of the International Conference on Computational Linguistics (COLING 2012), pp. 895–910 (2012)
25. Fox, L.M., Williams, L.A., Hunter, L., Roeder, C.: Negotiating a text mining license for faculty researchers. Information Technology and Libraries **33**(3), 5–21 (2014)
26. Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A.: GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics **17**(Suppl. 1), S74–S82 (2001)
27. Gautama: Nyaaya Suutras (150 CE)

28. Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., Karen, O., Sarker, A., Smith, K., Gonzalez, G.: Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark. In: Evaluating Resources for Health and Biomedical Text Processing (BioTxtM2014). Reykjavik, Iceland (2014). URL http://www.nactem.ac.uk/biotxtm2014/programme.php
29. Golik, W., Warnier, P., Nédellec, C.: Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In: Proc. 9th Intl Conf. Terminology and Artificial Intelligence (TIA 2011), pp. 37–39 (2011)
30. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: COLING, vol. 96, pp. 466–471 (1996)
31. Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., Quintard, L.: Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In: Proceedings of the 5th Linguistic Annotation Workshop, pp. 92–100. Portland, Oregon, USA (2011). URL http://www.aclweb.org/anthology/W11-0411. Poster
32. Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics **45**(5), 885 – 892 (2012). DOI 10.1016/j.jbi.2012.04.008
33. Haverinen, K., Ginter, F., Laippala, V., Viljanen, T., Salakoski, T.: Dependency-based propbanking of clinical Finnish. In: Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV), pp. 137–141. ACL (2010)
34. Hersh, W., Kalpathy-Cramer, J., Müller, H.: The ImageCLEFmed medical image retrieval task test collection. Journal of Digital Imaging **22**, 648–655 (2009)
35. Hirschman, L., Robinson, P., Burger, J., Vilain, M.: Automating coreference: The role of annotated training data. In: Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, pp. 118–121 (1997)
36. Hripcsak, G., Rothschild, A.S.: Agreement, the F-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association **12**(3), 296–298 (2005)
37. Ide, N., Meyers, A., Pradhan, S., Tomanek, K. (eds.): Proceedings of the 5th Linguistic Annotation Workshop. Association for Computational Linguistics, Portland, Oregon, USA (2011). URL http://www.aclweb.org/anthology/W11-04
38. Ide, N., Xia, F. (eds.): Proceedings of the Sixth Linguistic Annotation Workshop. Association for Computational Linguistics, Jeju, Republic of Korea (2012). URL http://www.aclweb.org/anthology/W12-36
39. Kedzia, P., Piasecki, M., Maziarz, M., Marcińczuk, M.: Recognising compositionality of multi-word expressions in the wordnet oriented perspective. In: Advances in Artificial Intelligence and Its Applications, pp. 240–251. Springer (2013)
40. Kilicoglu, H., Rosemblat, G., Fiszman, M., Rindflesch, T.C.: Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics **12**(1), 486 (2011)
41. Kim, J.D.: A generalized LCS algorithm and its application to corpus alignment. In: Proceedings of the 6th International Joint Conference on Natural Language Processing, pp. 14–18 (2013)
42. Kim, J.D.: Sharing reference texts for interoperability of literature annotation. In: Proceedings of the 5th international symposium on languages in biology and medicine, pp. 57–61 (2013)
43. Kim, J.D., Ohta, T., Tateisi, Y., Mima, H., Tsujii, J.: XML-based linguistic annotation of corpus. In: Proceedings of The First NLP and XML Workshop, pp. 47–53 (2001)
44. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics **19**(Suppl. 1), 180–182 (2003)
45. Kim, J.D., Wang, Y.: PubAnnotation: a persistent and sharable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pp. 202–205. Association for Computational Linguistics (2012)
46. Lee, H.J., Shim, S.H., Song, M.R., Lee, H., Park, J.C.: CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations. BMC Bioinformatics **14**(1), 323 (2013)

47. Levin, L., Stede, M. (eds.): Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (2014). URL http://www.aclweb.org/anthology/W14-49
48. Lin, J.: Is searching full text more effective than searching abstracts? BMC Bioinformatics **10**(46) (2009)
49. Lu, Z., Kao, H.Y., Wei, C.H., Huang, M., Liu, J., Kuo, C.J., Hsu, C.N., Tsai, R.T., Dai, H.J., Okazaki, N., et al.: The gene normalization task in BioCreative III. BMC Bioinformatics **12**(Suppl 8), S2 (2011)
50. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics **19**(2), 313–330 (1993)
51. McIntosh, T., Curran, J.R.: Challenges for automatically extracting molecular interactions from full-text articles. BMC Bioinformatics **10**(311) (2009)
52. Mihăilă, C., Ohta, T., Pyysalo, S., Ananiadou, S.: BioCause: Annotating and analysing causality in the biomedical domain. BMC bioinformatics **14**(1), 2 (2013)
53. Mitchell, A., Strassel, S., Huang, S., Zakhary, R.: Ace 2004 multilingual training corpus. Linguistic Data Consortium, Philadelphia (2005)
54. Molla, D., Santiago-Martinez, M.E.: Development of a corpus for evidence based medicine summarisation. In: Proceedings of the Australasian Language Technology Association Workshop, pp. 86–94 (2011)
55. Morgan, A.A., Hirschman, L., Colosimo, M., Yeh, A.S., Colombe, J.B.: Gene name identification and normalization using a model organism database. J. Biomedical Informatics **37**(6), 396–410 (2004). DOI 10.1016/j.jbi.2004.08.010. URL http://dx.doi.org/10.1016/j.jbi.2004.08.010
56. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., et al.: Overview of BioCreative II gene normalization. Genome Biology **9**(Suppl 2), S3 (2008)
57. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The Quaero French Medical Corpus: A resource for medical entity recognition and normalization. In: Fourth workshop on building and evaluating resources for health and biomedical text processing (2014)
58. Neves, M.: An analysis on the entity annotations in biological corpora. F100 Research **3(96)** (2014)
59. Nobata, C., Dobson, P.D., Iqbal, S.A., Mendes, P., Tsujii, J., Kell, D.B., Ananiadou, S.: Mining metabolites: extracting the yeast metabolome from the literature. Metabolomics **7**(1), 94–101 (2011)
60. Nunes, T., Campos, D., Matos, S., Oliveira, J.L.: BeCAS: biomedical concept recognition services and visualization. Bioinformatics **29**, 1915–1916 (2013)
61. Ogren, P.: Knowtator: a Protege plugin for annotated corpus construction. In: HLT-NAACL 2006 Companion Volume (2006a)
62. Ogren, P.: Knowtator: a plug-in for creating training and evaluation data sets for biomedical natural language systems. In: The International Protege conference, pp. 73–76 (2006b)
63. Ohta, T., Kim, J.D., Pyysalo, S., Wang, Y., Tsujii, J.: Incorporating GENETAG-style annotation to GENIA corpus. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 106–107. Association for Computational Linguistics (2009)
64. Ohta, T., Pyysalo, S., Tsujii, J., Ananiadou, S.: Open-domain anatomical entity mention detection. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, pp. 27–36. Association for Computational Linguistics (2012)
65. Ohta, T., Tateisi, Y., Kim, J.D., Mima, H., Tsujii, J.: The GENIA corpus: an annotated corpus in molecular biology. In: Proceedings of the Human Language Technology conference (2002)
66. Pareja-Lora, A., Liakata, M., Dipper, S. (eds.): Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Association for Computational Linguistics, Sofia, Bulgaria (2013). URL http://www.aclweb.org/anthology/W13-23

67. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Morante, R.: QA4MRE 2011-2013: Overview of question answering for machine reading evaluation. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 303–320. Springer (2013)
68. Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., Savova, G.: Task 1: ShARe/CLEF eHealth evaluation lab 2013. Online Working Notes of CLEF, CLEF **230** (2013)
69. Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W.W., Savova, G.: Evaluating the state of the art in disorder recognition and normalization of the clinical narrative
70. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–27. Association for Computational Linguistics (2011)
71. Pradhan, S.S., Ramshaw, L., Weischedel, R., MacBride, J., Micciulla, L.: Unrestricted coreference: Identifying entities and events in OntoNotes. In: Semantic Computing, 2007. ICSC 2007. International Conference on, pp. 446–453. IEEE (2007)
72. Prasad, R., McRoy, S., Frid, N., Joshi, A., Yu, H.: The biomedical discourse relation bank. BMC BioInformatics **12**(88) (2011)
73. Pustejovsky, J., Stubbs, A.: Natural language annotation for machine learning. O'Reilly Media (2012)
74. Pyysalo, S., Ananiadou, S.: Anatomical entity mention recognition at literature scale. Bioinformatics (2013)
75. Pyysalo, S., Ohta, T., Miwa, M., Cho, H.C., Tsujii, J., Ananiadou, S.: Event extraction across multiple levels of biological organization. Bioinformatics **28**(18), i575–i581 (2012)
76. Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J., Ananiadou, S.: Overview of the infectious diseases (ID) task of BioNLP Shared Task 2011. In: Proceedings of the BioNLP Shared Task 2011 Workshop, pp. 26–35. Association for Computational Linguistics (2011)
77. Raghavan, P., Fosler-Lussier, E., Lai, A.M.: Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. In: AMIA Annual Symposium Proceedings, vol. 2012, p. 1366. American Medical Informatics Association (2012)
78. Ramanan, S., Nathan, P.S.: Adapting Cocoa, a multi-class entity detector, for the CHEMD-NER task of BioCreative IV (2013)
79. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A.: Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics **42**(5), 950–66 (2009)
80. Roberts, K., Harabagiu, S.M., Skinner, M.A.: Structuring Operative Notes using Active Learning. In: Proceedings of the 2014 BioNLP Workshop, pp. 68–76 (2014)
81. Roberts, K., Masterton, K., Fiszman, M., Kilicoglu, H., Demner-Fushman, D.: Annotating question decomposition on complex medical questions. In: Language Resources and Evaluation Conference (2014)
82. Roberts, K., Masterton, K., Fiszman, M., Kilicoglu, H., Demner-Fushman, D.: Annotating Question Types for Consumer Health Questions. In: Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (2014)
83. Savova Guergana; Pradhan, S.P.M.S.W.C.W.E.N.: Annotating the clinical text - MiPACQ, ShARe, SHARPn and THYME corpora. In: N. Ide, J. Pustejovsky (eds.) This volume. Springer (2015)
84. Shah, P.K., Perez-Iratxeta, C., Bork, P., Andrade, M.A.: Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics **4**(1) (2003). DOI 10.1186/1471-2105-4-20. URL http://dx.doi.org/10.1186/1471-2105-4-20
85. Smith, B., Ceusters, W.: Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied ontology **5**(3), 139–188 (2010)

86. Stede, M., Huang, C.R., Ide, N., Meyers, A. (eds.): Proceedings of the Third Linguistic Annotation Workshop. Association for Computational Linguistics, Suntec, Singapore (2009). URL http://www.aclweb.org/anthology/W09-30
87. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 102–107. Association for Computational Linguistics (2012)
88. Stubbs, A.: A methodology for using professional knowledge in corpus annotation. Ph.D. thesis, Brandeis University (2013)
89. Stubbs, A., Uzuner, O.: De-identification of medical records through annotation. In: N. Ide, J. Pustejovsky (eds.) Handbook of Linguistic Annotation. Springer (2015)
90. Tanabe, L., Wilbur, W.J.: Tagging gene and protein names in full text articles. In: Natural language processing in the biomedical domain, pp. 9–13 (2002)
91. Tateisi, Y., Yakushiji, A., Ohta, T., Tsujii, J.: Syntax annotation for the GENIA corpus. In: Second international joint conference on natural language processing: Companion volume, pp. 220–225 (2005)
92. Temnikova, I.P., Cohen, K.B.: Recognizing sublanguages in scientific journal articles through closure properties. In: Proceedings of BioNLP 2013 (2013)
93. Thompson, P., Iqbal, S.A., McNaught, J., Ananiadou, S.: Construction of an annotated corpus to support biomedical information extraction. BMC Bioinformatics **10**(1), 349 (2009)
94. Thompson, P., Nawaz, R., McNaught, J., Ananiadou, S.: Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics **12**(1), 393 (2011)
95. Van Auken, K., Schaeffer, M.L., McQuilton, P., Laulederkind, S.J., Li, D., Wang, S.J., Hayman, G.T., Tweedie, S., Arighi, C.N., Done, J., et al.: BC4GO: a full-text corpus for the BioCreative IV GO task. Database **2014**
96. Van Mulligen, E.M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J.A., Furlong, L.I.: The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. Journal of Biomedical Informatics **45**(5), 879–884 (2012)
97. Verspoor, K., Cohen, K.B., Hunter, L.: The textual characteristics of traditional and Open Access scientific journals are similar. BMC Bioinformatics **10** (2009)
98. Verspoor, K., Cohen, K.B., Lanfranchi, A., Warner, C., Johnson, H.L., Roeder, C., Choi, J.D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Jr., W.A.B., Bada, M., Palmer, M., Hunter, L.E.: A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics **13**(207) (2012)
99. Verspoor, K., Yepes, A.J., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., Plazzer, J.P.: Annotating the biomedical literature for the human variome. Database: The Journal of Biological Databases and Curation (2013)
100. Xue, N., Poesio, M. (eds.): Proceedings of the Fourth Linguistic Annotation Workshop. Association for Computational Linguistics, Uppsala, Sweden (2010). URL http://www.aclweb.org/anthology/W10-18