Semantic Annotations for Biology — A Corpus Development Initiative at the Jena University Language & Information Engineering (JULIE) Lab

U. Hahn, E. Beisswanger, E. Buyko, M. Poprat, K. Tomanek, J. Wermter

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Fürstengraben 30, D-07743 Jena, Germany
{hahn|beisswanger|buyko|poprat|tomanek|wermter}@coling-uni-jena.de

Abstract

We provide an overview of corpus building efforts at the Jena University Language & Information Engineering (JULIE) Lab, which are focused on life science documents. Special emphasis is laid on semantic annotations in terms of a large amount of biomedical named entities (almost 100 entity types), semantic relations, as well as discourse phenomena, reference relations in particular.

1. Introduction

The construction of natural language corpora annotated with various types of linguistic meta data (typically ranging from POS tags over syntactic structure information up to named entities, semantic relations, and discourse structures) has been one of the true success stories for NLP. The PENN TREEBANK (Marcus et al., 1993) and the PENN PROPBANK (Palmer et al., 2005), e.g., have become a *de facto* standard for the coverage of the general newspaper language of English (though with a slant towards the economic domain). Based on the human-supplied annotations they contain, a variety of supervised machine learning algorithms have been trained on this meta data to subsequently operate on unlimited amounts of unseen data at different rates of accuracy.

Soon it turned out that shifts in the genre and the domain (not to mention language shifts) were considered harmful for the performance of taggers, parsers, named entity recognizers, etc. trained on newspaper annotations. In effect, their performance degraded significantly when ported to domains such as biology and medicine, or to genres such as scientific articles or abstracts (Hahn and Wermter, 2004).

In the life sciences, this observation gave rise to the development of domain-specific corpora — GENIA being the first, most prominent example of these efforts (Kim et al., 2003) (approximately 500,000 tokens annotated with several entity types), and PENNBIOIE (Kulick et al., 2004) (approximately 500,000 tokens with 22 distinct entity types) being another example built with more rigor and taking into consideration some of the shortcomings of GENIA. Still, both corpora are limited in scope because GENIA deals with transcription factors of human blood cells, while PENNBIOIE deals with oncology and CYP450 proteins only. Again, it was shown that porting named entity recognizers trained on GENIA or PENNBIOIE data underperformed in biological fields other than the ones already covered (Tomanek et al., 2007b).

Obviously, new fields (in the life sciences) not already covered by existing annotated corpora need new annotations unless one is willing to pay a high price, cashing in the performance-degrading effects of domain change with GENIA- or PENNBIOIE-trained ML systems. Since the provision of new annotations is rather costly in terms of acquiring, training and supervising (life science) expert staff, our solution was to develop an annotation methodology based on Active Learning (Tomanek et al., 2007a). It turned out that applying this approach to the corpus annotation task, manpower expenses could be lowered by up to 75% for entity annotations, while keeping almost the same level of annotation quality. Given such a methodology, it now seems more feasible to go for (sub)domain changes without committing to overly excessive annotation costs.

In this paper, we report on our activities to set up two new life science corpora — one for the domain of gene regulation and expression (of E. coli), the other dealing with immunogenetics. This corpus construction initiative is embedded in two major projects our lab is involved in, *viz.* the BOOTSTREP[1] and the STEMNET[2] project, respectively. Both projects deal with the development of information retrieval and information extraction systems for the life sciences.

While low-level text formatting and syntactic annotations are contained in these corpora as well, we here focus on semantic issues in that we distinguish between annotation efforts dealing with named entities, relations, and discourse phenomena in terms of anaphora. While our annotations for named entities make use of Active Learning, the annotations of relations as well as referential discourse structures have been performed in the classical manner where a couple of relevant abstracts were randomly selected for annotation.

2. Semantic Annotation

In the following, we will shortly describe the annotation environment used for developing the two corpora and then focus on our biological corpus annotations which deal with named entities, relations and events, as well as referential discourse phenomena.

[1] www.bootstrep.eu

[2] www.stemnet.de

2.1. Annotation Environment

All our annotations were performed using the Jena ANnotation Environment (JANE) (Tomanek et al., 2007b). It supports the whole annotation workflow, including the set-up of annotation projects, user management, annotation itself via an external editor, monitoring of the annotation process, and deployment of the annotated material. In our annotation routines, several annotators were involved. Thus, JANE's user management had to take care that annotators work only on those annotation projects they were assigned to. We also consider the central storage of all annotation-relevant data in a database as beneficial to keep control of the distributed annotation efforts.

The most striking innovation behind JANE is its support for Active Learning (AL) to speed up the annotation process with no loss of annotation quality. AL is an intelligent sampling strategy which selects in an iterative manner those examples (in our case, sentences) for manual annotation which are expected to be the most informative for classifier learning. JANE employs a committee-based AL approach (Tomanek et al., 2007a) where in every iteration an ensemble of classifiers is trained on the already annotated material. Each of these classifiers make a prediction on the unlabeled examples. Examples on which the committee's members highly disagree are considered informative and thus are directed to human annotators for providing further labeling advice.

We have extensively employed JANE's AL facilities during our entity annotation cycles since this annotation data mainly serves as training material for machine-learning-based entity taggers. Using AL we could thus considerably boost the annotation rate and density (in terms of entity mentions contained in the annotated sentences) without sacrificing the quality of this meta data.

2.2. Named Entities

Entities are at the heart of any approach to develop semantic search and information extraction systems. Hence, without their proper recognition it is usually impossible to identify and extract the semantic information locked in natural language text. Thus far, our annotations encompass 10 entity categories with 97 distinct entity types (see Table 1). We measured interannotator agreement for three entity categories ('cytokine and growth factor receptors', 'organisms and organism attributes', and 'transcription regulators and ligands') in terms of the "authoritative annotator" F-score (Kim and Tsujii, 2006). Given two human annotators, this interannotator agreement metrics basically frames the annotations of one annotator to be the "gold standard" against which the other annotator's annotations are evaluated. For the three categories we got interannotator F-scores of 80.2, 85.1, and 65.0, respectively. The reason why the score for the first entity category is lower than the second one might be that the first one is a specific protein function which is rather difficult to annotate. The F-score for the third category, down at 65.0, rises to 80.0, when not only exact matches but also overlapping matches are accepted. This indicates that finding the right boundaries of entity mentions in texts was a particular challenge for the annotators of transcription regulators and ligands.

For each entity category, a large collection of scientific abstracts were selected from MEDLINE,[3] a huge international literature database that covers much of the literature in medicine and biology, using a MESH[4] query.[5] From these abstracts, single sentences were selected by JANE's Active Learning component for manual annotation.

Table 1 gives an overview of the semantic types and amount of entities being annotated. As can be seen, so far we were able to annotate 97 distinct immunogenetic and gene regulation entity types, ranging from various specific protein types (cytokines, growth factors and their receptors, major and minor histocompatibilty antigens, transcription regulators) and chemicals (ligands) over various kinds of immune cells (various sorts of T-, B-, NK- and dendritic cells) and blood progenitor cells to organisms and genomic variations.

The guidelines for these annotations were developed in iterations, with annotators first annotating a set of "guideline consolidation" texts with many possible entity mentions of the types to be annotated.[6] In general, the guidelines stated to be very strict regarding the inclusion of pre- or postmodifiers of entity terms and only include those which typified (i.e., characterized distinguishably) the entity under consideration, while those were disregarded which merely described (i.e., added additional information to) the entity. In cases when this was not clear, the modifiers were to be ignored. Guidelines to specific entity categories were then interactively refined until they were deemed to be stable for consistent annotations. Once this was attained, the annotation effort was continued via Active Learning in JANE (see Subsection 2.1.).

With respect to the entities being linked and normalized to their respective objects in the (biological) world, we ensured that all immune and progenitors cell entity types are anchored to concepts from the OBO Cell Type Ontology.[7] Concerning the proteins, we aimed at linking as many of them as possible to concepts from the (functional branch) of the Gene Ontology (GO)[8] and to the INOH Molecule Role Ontology.[9] This, however, was not always possible, either due to the high abstraction level of GO concepts (which makes them hard to link to text entity mentions) or due to conceptual gaps in both ontologies. This is one of the rare instances where annotations are guided and influenced by a dedicated ontology.[10]

[3] `http://www.ncbi.nlm.nih.gov/sites/entrez`

[4] The Medical Subject Headings (MESH, `http://www.nlm.nih.gov/mesh`) is a high-coverage controlled vocabulary created and used for indexing the MEDLINE database by the United States National Library of Medicine (NLM).

[5] This query basically consisted of the MESH terms equivalent to the entity categories in Table 1.

[6] These texts were selected via matching an entity dictionary against a collection of MEDLINE abstracts and then selecting those with the highest number of matches.

[7] `http://obofoundry.org/cgi-bin/detail.cgi?id=cell`

[8] `http://www.geneontology.org`

[9] `http://obofoundry.org/cgi-bin/detail.cgi?id=molecule_role`

[10] Currently, we are also evaluating the performance of automatically linking annotated proteins to their respective biologically normalized database entries in UNIPROT (`www.uniprot.`

domain	entity category	# of distinct entity types	# of tokens annotated
immunogenetics	cytokines and growth factors	7	276,570
immunogenetics	cytokine and growth factor receptors	7	223,314
immunogenetics	antigens (e.g. CD antigens)	6	196,063
immunogenetics	minor histocompatibility antigens	6	187,496
immunogenetics	organisms and organism attributes	18	173,943
immunogenetics	T cells and natural killer cells	15	158,856
immunogenetics	B cells and dendritic cells	14	164,790
immunogenetics	hematopoietic progenitor cells	10	146,482
immunogenetics	genomic variations	6	139,961
gene regulation	transcription regulators and ligands	8	203,900
	Σ	97	$>$ 2,011,336

Table 1: Large-scale and entity-rich semantic annotations for immunogenetics and gene regulation

2.3. Semantic Relations

Besides information about named entities mentioned in the texts, we are interested in extracting semantic relations which hold between the named entities. Obviously, being able to drill down named entities is a prerequisite for relation extraction. The annotation of semantic relations between entities is already in the focus of many annotation initiatives. This holds for newspaper corpora such as MUC-7 (MUC-7, 1998) or ACE (Doddington et al., 2004), but also for life science corpora such as BIOINFER (Pyysalo et al., 2007) or BIOCREATIVE II's protein-protein interaction corpus (Hirschman et al., 2007).

We aim to provide a corpus suited for the extraction of semantic relations between entities in the biomedical domain, especially for the domain of gene regulation. While literal relation mentions (as in *"X regulates Y"* or *"X binds Y"*) still occur, the majority of relation mentions are described in a much more complex way. For example, the sentence *"Deletion of the arcA gene caused about a 2-fold increase in the ptsG expression"* contains the description of negative regulation of ptsG expression by the arcA gene. This relation, however, can only be inferred from the following statements mentioned in the text above by a reader with a sound biomedical background: *"deletion of arcA"* and *"increase in the ptsG expression"*. Proper detection of these statements is crucial for the identification of the semantic relation between the *'arcA'* and the *'ptsG'* gene. Given such complex expressions of semantic relations in our domain of gene regulation, we aim to provide annotations in such a way that they are also helpful for the automatic detection of complex and indirectly mentioned gene regulation relations. The corpus contains thus two general types of annotations:

- Annotations of relations between entities

  Each semantic relation is defined as an ordered binary relation between named entities occurring in the same sentence.

- Annotations of relation triggers

  Relation triggers are text spans (mostly single words) which refer to statements relevant to the detection of semantic relation of interest. Verbs and their normalizations constitute the major part of relation triggers. In our example sentence, *'deletion'*, *'increase'* and *'expression'* are the relation triggers.

Currently, our annotations cover the domain of the regulation of gene expression in *E. coli*, only. We focus particularly on interactions between proteins in the process of gene regulation. Here, the key players are genes (proteins) that can participate in the relation `regulation_Of_Gene_Expression (arg1, arg2)` where `arg1` is the regulatory gene (or agent) and `arg2` is the regulated gene (patient or participant). This relation can be enriched with further, biologically relevant, information, e.g., about the `polarity` of the relation (`positive`, `negative` and `unspecified`), or information about the `physical_contact` between the actors (`arg1` and `arg2`) involved (which can be `true` (direct binding), `false` (no direct binding) or `unspecified`). Thus, for our example sentence we instantiate the relation `regulation_Of_Gene_Expression (arcA, ptsG)`, with attributes `polarity = negative` and `physical_contact = unspecified`.

For the annotation of gene regulation relations, we selected three types of relation triggers that may occur when the relation of interest is dealt with in the text: First, `gene_expression` is defined as the mention of the process by which a coding sequence of a gene is converted into a mature gene product (or products). Second, `gene_regulation` is defined as the mention of any process that refers to the modulation of the frequency, rate or extent of gene expression. Finally, `experimental_intervention` is defined as the mention of any process of experimental genetic modifications. Accordingly, we annotate our example sentence as shown in Figure 1.

The annotated corpus contains currently 309 PubMed[11] abstracts which boils down to 3,140 sentences and approximately 86,000 tokens.

`org`) and ENTREZ GENE (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`).

[11] `http://www.pubmed.org`

```
<regulation_Of_Gene_Expression  polarity = negative physical_contact = unspecified>
        <experimental_intervention>Deletion</experimental_itervention>
        of the arcA gene caused about a 2-fold
        <gene_regulation>increase</gene_regulation>
        in the ptsG
        <gene_expression>gene expression</gene_expression>
</regulation_Of_Gene_Expression>
```

Figure 1: Annotation Example for the Semantic Relation `regulation_Of_Gene_Expression`

2.4. Referential Expressions

The need to cope with referential expressions (coreferences, basically) for information extraction (IE) has been acknowledged and embedded as a special task in recent IE challenges such as MUC-6 (MUC-6, 1995), MUC-7 (MUC-7, 1998) or ACE2005 (Doddington et al., 2004)). Also for an information extraction system for court opinions, Al-Kofahi et al. (1999), for instance, show that the resolution of referential expressions has a positive impact on the overall system performance.

In the life sciences domain, however, in none of the bio-NLP challenges, up until now, coreference resolution has been considered as (part of) a task. Consequently, there are no MUC- or ACE-comparable life sciences corpora including reference annotations. Only few corpora, e.g., MEDSTRACT (Castaño et al., 2002; Pustejovsky et al., 2007), BIOINFER (Pyysalo et al., 2007) and GENIA,[12] have already extended their annotations to incorporate reference expressions as well. As a drawback, however, in these already existing corpora only few documents were annotated. Annotating corpora with referential expressions is a complex activity because both linguistic expertise and domain knowledge is required for a sound annotation. Existing MUC or ACE annotations guidelines cannot be simply transferred without taking into consideration the particularities of the biomedical sublanguage. Furthermore, referential expressions annotated in existing biomedical corpora are limited to coreference annotations (disregarding other relations such as `part-of` or `derives-from`, see Poprat and Hahn (2007a)). In addition, some difficulties with respect to the annotation boundaries and the influence of linguistic modifiers have not been sufficiently addressed yet.

Therefore, we decided to define more comprehensive guidelines and to annotate referential expressions in biomedical abstracts in a more detailed manner (Poprat and Hahn, 2007b). Basically, our guidelines define as annotations mostly the heads of base nominal phrases (NPs). However, in some complex base NPs, we also have to annotate parts of base NPs (e.g., "*IL-2-dependent*" – "*IL2*") in particular when these expressions can be paraphrased (here: "*dependent on IL-2*"). Furthermore, our guidelines also allow to annotate modifiers (mainly prepositional phrases and adjectives), if necessary. This is the case when the head of the NP is modified in order to distinguish it from other (monotonous) heads. For instance, in the mentions "*cells from mouse*" and "*cells from zebrafish*" occurring in the same text, "*cells*" are further distinguished by the organism "*mouse*" and "*zebrafish*", respectively. In the example "*CD34(+) cells*" vs. "*purified CD34(+) cells*" vs. "*irradiated CD34(+) cells*", the common head "*CD34(+) cells*" is modified by the adjectives "*purified*" and "*irradiated*". Taking modifiers into account influences the decision whether a linguistic entity is still coreferential or not. Finally, in our guidelines we also have (currently informal) rules that define the annotation of relations between referential expressions.

For now, our corpora comprise 56 abstracts, with 21,530 tokens. The annotations were carried out by a graduate biologist trained in a basic understanding of linguistic notions (such as what constitutes a modifier, a base noun phrase, etc.). In these documents we found 1,178 coreferential expressions (954 repetitions, 107 pronominal and 117 nominal anaphora), 130 subgrouping expressions and 95 bridging anaphora (e.g., part-of relations). These numbers, at the first glance, may not be so shiny but are more than on a par with comparable efforts. In comparison, the first version of MEDSTRACT (Castaño et al., 2002) comprised 32 abstracts (about 6,000 tokens) with (only) 72 coreferences and 13 subgrouping anaphora. The second version of MEDSTRACT (still not publicly available) will contain around 1,300 coreference annotations from 370 abstracts (Pustejovsky et al., 2007). Though lot of work remains to be done (e.g., checking the quality of both the guidelines and the annotations), the annotations contained in our corpus are the most differentiated and, in quantitative terms, largest ones currently available for the biomedical domain.

3. Conclusion and Outlook

Empirical evidence indicates that genre and domain changes almost always dictate a need for new annotations, at the semantic level, in particular. The need comes with annotation costs in time, manpower, etc. In the JULIE Lab, we support our entity annotations with Active Learning – both in terms of annotation methodology and tool-based technology – which reduces the amounts of efforts for annotations significantly, while preserving, by and large, the same level of annotation quality.

Semantic meta data is annotated in two different corpora for two different domains in the life sciences, *viz.* gene regulation and expression, as well as immunogenetics, both areas not covered by any annotation activities before. These efforts reflect our involvement in two major biomedical text mining projects, BOOTSTREP and STEMNET, respectively. We deal with crucial named entities in these fields, with numbers of entity types (97 types) and fine conceptual granularity – reflecting challenging requirements of our task domain – unmatched by existing corpora. These entity annotations are currently complemented by relation and event annotations only recently provided in the GE-

[12] `http://nlp.i2r.a-star.edu.sg/medco.html`

NIA corpus (PENNBIOIE does not provide any relations). Finally, we augment entity and relation annotations by a wide range of referential discourse phenomena not found in any of the already available resources (although the GENIA team announced the distribution of reference annotations in the near future).

Our goal is to create a semantic meta data resource that compares to GENIA and PENNBIOIE in terms of quantitative coverage parameters but excels in terms of depth and annotation quality. We also consider our work as a test case for the feasibility of diverse and large-scale corpus annotations, once Active Learning is the annotation methodology of choice. This is a crucial goal because domain and genre shifts will almost always require (new) annotations. Hence, coping with this bottleneck is a prerequisite for flexible adaptation of HLT tools (not only) in the life sciences field. The real benefit of all this annotation work, however, will only become evident when sufficiently reliable inter-annotator agreement values are communicated and system modules trained on this meta data are integrated into a common architecture to deliver sophisticated text analytics.

Acknowledgements

This research was funded by the EC within the BOOTSTREP project (FP6-028099) and by the German Ministry of Education and Research within the STEMNET project (01DS001A-C).

4. References

Khalid Al-Kofahi, Brian Grom, and Peter Jackson. 1999. Anaphora resolution in the extraction of treatment history language from court opinions by partial parsing. In *ICAIL '99 – Proceedings of the 7th International Conference on Artificial Intelligence and Law*, pages 138–146. Oslo, Norway, 14-17 June, 1999.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution for Natural Language Processing*. Alicante, Spain, June 3-4, 2002.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program: Tasks, data, & evaluation. In *LREC 2004 – Proceedings of the 4th International Conference on Language Resources and Evaluation. Vol. 3*, pages 837–840. Lisbon, Portugal, 26-28 May 2004.

Udo Hahn and Joachim Wermter. 2004. High-performance tagging on medical texts. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, volume 2, pages 973–979. Geneva, Switzerland, August 23-27, 2004.

Lynette Hirschman, Martin Krallinger, and Alfonso Valencia, editors. 2007. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid: CNIO Centro Nacional de Investigaciones Oncológicas.

Jin-Dong Kim and Jun'ichi Tsujii. 2006. Corpora and their annotation. In Sophia Ananiadou and John McNaught, editors, *Text Mining for Biology and Biomedicine*, pages 179–211. Norwood, MA: Artech House.

Jin-Dong Kim, Tomoka Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2003. GENIA corpus: A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):i180–2

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of the HLT-NAACL 2004 Workshop 'Linking Biological Literature, Ontologies and Databases: Tools for Users – BioLink 2004'*, pages 61–68. Boston, MA, USA, May 2004.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The PENN TREEBANK. *Computational Linguistics*, 19(2):313–330.

MUC-6. 1995. *Proceedings of the 6th Message Understanding Conference*. Columbia, Maryland, November 6-8, 1995. San Mateo, CA: Morgan Kaufmann.

MUC-7. 1998. *Proceedings of the 7th Message Understanding Conference, NYU.*

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Michael Poprat and Udo Hahn. 2007a. An investigation into the reusability of biomedical terminologies for the resolution of referential expressions. In *BioLINK 2007 – Proceedings of the BioLINK SIG 2007. The Annual Meeting of the ISMB BioLINK Special Interest Group on Text Data Mining, in Association with ISMB 2007*, pages 39–42. Vienna, Austria, July 19, 2007.

Michael Poprat and Udo Hahn. 2007b. Quantitative data on referring expressions in biomedical abstracts. In *BioNLP at ACL 2007 – Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing*, pages 193–194. Prague, Czech Republic, June 29, 2007.

James Pustejovsky, José Castaño, Maciej Kotecki, and Brent Cochran. 2007. An annotated biological corpus. In *BioLINK 2007 – Proceedings of the BioLINK SIG 2007. The Annual Meeting of the ISMB BioLINK Special Interest Group on Text Data Mining*, pages 51–54. Vienna, Austria, July 19, 2007.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *Bioinformatics*, 8(50).

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007a. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *EMNLP-CoNLL 2007 – Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 486–495. 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.