Active Learning-Based Corpus Annotation — The PATHOJEN Experience

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Erik Faessler
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Fürstengraben 30, D-07743 Jena, Germany
udo.hahn@uni-jena.de

Abstract

We report on basic design decisions and novel annotation procedures underlying the development of PATHOJEN*, a corpus of* MEDLINE *abstracts annotated for pathological phenomena, including diseases as a proper subclass. This named entity type is known to be hard to delineate and capture by annotation guidelines. We here propose a two-category encoding schema where we distinguish short from long mention spans, the first covering standardized terminology (e.g. diseases), the latter accounting for less structured descriptive statements about norm-deviant states, as well as criteria and observations that might signal pathologies. The second design decision relates to the way annotation instances are sampled. Here we subscribe to an Active Learning-based approach which is known to save annotation costs without sacrificing annotation quality by means of a sample bias. By design, Active Learning picks up 'hard' to annotate instances for human annotators, whereas 'easier' ones are passed over to the automatic classifier whose models already incorporate and gradually improve with previous annotation experience.*

1 Introduction

Biomedical language processing, its (supervised) machine learning-based branch, in particular, has become a resource-greedy enterprise in the sense that for virtually each level of linguistic processing — such as sentence and token splitting, POS tagging, parsing, named entity recognition and relation extraction — human-made or human-monitored meta data have to be assigned to the underlying raw text. In the meantime, we have entered a phase in which semantic, rather than syntactic considerations play the most important role for annotation activities.

This development is evidenced, for instance, by the flourishing development of several phenotype corpora (see Section 2 for a detailed discussion). They have been annotated with semantic meta data covering a wide range of pathological phenomena, diseases in particular, but also including treatments, tests and therapies, as well as drugs and their relations to diseases. From a genre perspective, these efforts not only incorporate standard MEDLINE abstracts but increasingly include clinical notes and electronic patient records. In these corpora, we witness a large variety and granular depth of named entity and relation types. At the same time, comparatively low values from inter-annotator agreement measurements are reported that further vary considerably depending on the entity and relation types being considered. Since this observation holds for virtually all reported annotation studies, it might signal the intrinsic complexity of annotations dealing with a broad range of pathological phenomena, including but not equaling diseases.

Indeed, pathological phenomena are obviously an instance of what we call *semantically sloppy* named entity types,[1] i.e. ones which are hard to delineate and capture by annotation guidelines. As a way to get out of the annotation dilemma for sloppy named entities, we propose a special encoding schema. It is based on the distinction between short and long mention spans where short mention spans are primarily coded for standard terminology (e.g. disease names), whereas long(er) mention spans are coded for descriptive verbal statements related to norm-deviant states and other quasi-pathological or pathology-signaling criteria and observations.

Irrespective of strictness or sloppiness of annotation targets, all annotation efforts suffer from enormous investments in (training and coding) time and the people involved in annotation projects. These considerations have sparked the idea to find alternatives to cost-intensive annotation procedures. One of these cost-cutting approaches uses *Active Learning* (AL).[2] It is based on the idea that the training of a classifier can be speeded up considerably when only particularly 'hard' annotation instances are presented to a human oracle for classification whereas the 'easy' ones are immediately delegated to a machine classifier (which disposes of a iteratively improved decision model based on the previous annotation experience). This sampling bias is crucial for AL and is counter to common random sampling.

In this paper, we apply AL, for the first time ever, to a larger scale annotation project the thematic scope of which are pathological phenomena in the broad sense from above. While dealing with a semantically sloppy named entity type, the annotation of pathological phenomena will be based on the short *vs.* long mention span annotation paradigm. Both design decisions are expected to yield a thematically richly, comparatively more consistently annotated corpus of pathological phenomena than previously attainable.

2 Related Work

Perhaps the most ambitious annotation enterprise, up until now, dealing with pathological phenomena resulted in the CLEF (Clinical E-Science Framework) corpus[3] 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 of clinical documents) which suggests that this fine-grained relationship annotation for clinical entities is a really hard task.[4] The latest round of the i2b2 Challenge[5] 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[6] or the Arizona Disease Corpus (AZDC).[7] Both only deal with *Disease* type annotations, a proper subset of *Pathological Phenomena* annotations. The EBI corpus contains 600 sentences from the Online Mendelian Inheritance in Man (OMIM) database,[1] 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* 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,[2] 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.*[8] report on a corpus which contains 1,556 annotations on 160 clinical notes using 658 unique concept codes from SNOMED-CT[3] 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.*[9] 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).[4]

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, for instance, locational and result-type relations). The short *versus* long form annotation we here advocate is meant as a compromise — rather than directly encoding pathology-relevant utterances with very detailed semantic meta data, we annotate relevant text spans that may be inspected for pathology information in more depth by suitable *Locus* and *Result* taggers, if specific applications require such information.

[1] http://www.ncbi.nlm.nih.gov/omim
[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T21
[3] http://www.ihtsdo.org/snomed-ct/
[4] http://www.cdc.gov/nchs/icd9.htm

In recent years, series of experiments have revealed that the sampling bias introduced by *Active Learning*[2] is more efficient for document annotation than standard random sampling, with almost no costs in effectiveness (annotation quality). At the core of this method is the idea that hard to decide annotation instances are preferably rendered to an oracle (the human annotator), whereas easier to decide annotation instances are directly handled automatically by the iteratively refined classifier. From a classifier perspective, AL thus identifies unlabeled samples that are highly "informative" (i.e. likely to increase classifier performance) for the classifier, avoiding non-informative samples (with no additional impact on classifier performance). Altogether, this yields high accuracy with a smaller training set size compared with random sampling.

This principle has already been investigated by Tomanek et al.[10] for named entity types which are prevailing in newspaper documents, such as *Persons*, *Organizations*, etc., as well as *Genes* and *Proteins* for the biological domain. They showed that, for both genres, cost reductions ranging from 48% to 72% in the number of tokens could be achieved, without seriously sacrificing annotation quality. In a series of follow-up studies these findings could be hardened with respect to multiple learning tasks (e.g. alternating between named entity recognition and syntactic parsing),[11] class imbalance (when named entity types occur at highly skewed frequency rates in a corpus),[12] or alternative cost models (e.g. involving annotation time as an empirically much more adequate efficiency criterion than the number of tokens being annotated).[13]

Experiments with biomedical named entities generated further empirical support for this approach in the life sciences domain. Perhaps the first study to apply AL methods to a biomedical task is due to Liu[14]. He applied AL to gene expression profiles of colon cancer, lung cancer, and prostate cancer samples and compared classification performance with that of random sampling. The results showed that employing AL can achieve high accuracy and significantly reduce the need for labeled training instances, thus outperforming passive learning. For example, to achieve 96% of the total positives, only 31 labeled examples were needed in AL, whereas in passive learning 174 labeled examples were required, a reduction rate of over 82%. In AL the areas under the receiver operating characteristic (ROC) curves were over 0.81, while in passive learning (random sampling) the areas under the ROC curves were below 0.50.

Kim et al.[15] exploited AL to compile a small but effective amount of training data for a biomedical named entity recognition system. In their experiments they showed that compared to random sampling the AL based approach (selecting only the most informative sentences from a given corpus for human annotation) helped reduce the human annotation effort significantly. Similar results were achieved by Tsuruoka et al.[16] with an AL-like annotation approach. The authors emphasize that cost savings are particularly high when the target named entities are sparse (see also[12]). Indeed, the minority class problem, where the number of exemplars from the non-target class substantially outnumber target class exemplars, is particularly important for life science documents. The harmful disproportion is usually dealt with employing class balancing methods (an application to digitized prostate histopathology annotation is described by Doyle et al.[17]).

For a clinical text classification task, Chen et al.[18] determine the assertion status of clinical concepts. The authors report results using the global Area under the average Learning Curve (ALC) score indicating that when the same number of annotated samples was used, AL strategies could generate better classification models (best ALC-0.77) than the passive learning method (random sampling) (ALC-0.74). Moreover, to achieve the same classification performance, AL strategies needed fewer samples than the random sampling method. All these results in different applications areas (see also work on proteomics data with focus on human protein-protein interaction prediction)[19] point consistently into the direction that AL annotation is truly beneficial for optimizing annotation processes. Based on these repeatedly generated findings, our annotation project uses AL as the annotation policy of choice.

3 Representation of Pathological Phenomena in PATHOJEN

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"*.

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 under no circumstances are considered as concrete diseases. Since the diagnostic tracing (determination or exclusion) of a disease, however, requires to refer to critical conditions and potentially indicative single observations of a patient, all these utterances 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.

To cope with both, concise mentions of pathological phenomena but also borderline cases, we decided to introduce two categories to be annotated, namely *short* and *long* mention spans. 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 a disease name), 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*, whereas 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 guidelines, a second annotation of type *long* has to be introduced that includes the *short* annotation and the anatomical specification *"in her left extremities"*. Type *long* annotations have various linguistic appearances. They occur, amongst others, as prepositional phrases (e.g. in *"[absence of auditory canals]$_{long}$"*), coordinations (e.g. in *"[markedly decreased serum IgG, IgA, and IgE levels]$_{long}$"*), and even entire sentences (e.g. in *"[The mass is compressing her trachea.]$_{long}$"*).

Our decision to annotate raw text along the dual *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. Basically, our annotation exercise, based on a two-way category system, provides a layered annotation where the *long* layer is open to further refinement. It allows researchers in follow-up studies to add further annotation details, reflecting their own specific pathological phenomena interests, by taking only *long* stretch annotations into account, instead of the whole abstract texts. Hence, to some extent, our coding strategy paves the way for the development of specialized *Pathological Phenomena* taggers.

4 Active Learning for Corpus Annotation

In standard annotation projects, instances to be annotated are usually sampled at random. In the *Active Learning* (AL) framework, however, the selection of examples which have to be manually annotated is intentionally biased such that the human labeling effort is minimized. This is achieved by selecting examples with (presumably) high utility for the classifier training. AL is thus a selective sampling technique where the learning protocol is in control of the data to be used. The goal of AL is to learn a good classifier with minimal human labeling effort. The class labels for examples which are considered most useful for the classifier training are queried iteratively from an oracle – typically a human annotator. In our scenario, the usefulness of examples is computed by a committee of classifiers as previously described by Tomanek et al. [10]

Algorithm 1 describes this selection routine in a generalized form. AL-based corpus annotation is an interactive process in which b sentences (at the very beginning, a reasonably chosen seed set) are selected by the AL engine for human annotation. Once the annotated data is supplied, the AL engine retrains its underlying classifier(s) on all available annotations and then re-classifies all unseen corpus items. After that, the most informative (i.e. deviant) b sentences from the set of newly classified data are selected for the next iteration round. In this approach, the time needed to select the next examples (which is the idle time of the human annotators) has to be kept at an acceptable limit of a few minutes only. The role of appropriate stopping criteria for the entire AL process – important to cash in the savings in effort – is discussed in Olsson and Tomanek [20].

Algorithm 1 Committee-based Active Learning

Given:

L: set of labeled examples, P: set of unlabeled examples

Algorithm:

loop until stopping condition is met

1. generate a committee C of classifiers $c_1, ..., c_k$: sample subset $L_i \subset L$ and train classifier c_i on it, $i \in [1, k]$
2. let each classifier c_i predict labels for all $e \in P$, yielding classifications $c_{e_1}, \ldots, c_{e_k}$
3. calculate disagreement $D_e(c_{e_1}, \ldots, c_{e_k})$ for all $e \in P$
4. select n examples $e \in P$ with highest D_e for annotation
5. move the annotated examples from P to L

To directly support the work of annotators in an AL scenario, we have developed JANE, the Jena ANnotation Environment.[21] It supports the whole annotation life-cycle, including the compilation of annotation projects, annotation itself (via an external editor), monitoring, and the deployment of annotated material. JANE consists of the *annotation repository* as central component, in which all annotation and project data are stored centrally, two *user interfaces* (namely one for the annotators and one for the administrator), and the *active learning* component, which interactively generates documents to speed up the annotation process. The components communicate with the annotation repository through a network socket – allowing JANE to be run in a distributed environment. JANE is largely platform-independent, because all components are implemented in Java.

JANE selects single, non-contiguous sentences from *different* documents. Since the context of these sentences is still crucial for many (semantic) annotation decisions, for each selected sentence its original context is added (though blocked from annotation). When AL selection has finished, a new document is compiled from these sentences (including their contexts) and uploaded to the annotation repository. The annotator can then proceed with annotation.

5 Active Learning-Supported Annotation Process

In this section, we describe the AL-supported annotation process for the PATHOJEN corpus based on the JANE annotation tool. To determine a reference document set for pathological phenomena we queried the MEDLINE 2011 baseline with "("Disease"[MeSH Terms] AND hasabstract[All Fields])" (i.e. documents with abstracts, indexed with the MeSH term "Disease" or a subordinated term). The extracted 65,397 documents (with 532k sentences) formed the *base corpus*. After an iterative process of annotator training and guideline revision (described in Hahn et al. [1]), we started the production phase for the PATHOJEN corpus using our AL machinery.

To illustrate the progress of AL-supported corpus development, we investigated the performance of JNET, our lab-internal named entity recognition (NER) machinery[22], trained on AL segments of the PATHOJEN corpus. 13 AL rounds were completed, each round contained a batch of 50 sentences ($b = 50$), presented within the context of the corresponding abstract. *Pathological Phenomena* annotations were provided according to the *short-long* category system described in Section 3. The complete corpus we used in the AL study comprises a set of 250 sentences (seed set) and 50 sentences from each of 13 annotation rounds. Thus, 900 sentences were manually annotated at the time of writing this paper. The corpus statistics are presented in Table 1.

Table 1: PATHOJEN corpus statistics.

Sentences	Tokens	*short*	*long*	Total annotated
900	13,007	1,292	707	1,999

For an evaluation study, we generated a test data set by randomly selecting 10% of the annotated sentences from the seed set and from each annotation round. Accordingly, altogether 90 common sentences were used in this evaluation study, 25 sentences extracted from the seed set and 5 sentences from each of the 13 annotation rounds. The test data sentences were removed from the training data. The remaining training sentences were used for training models of the JNET tagger, which were then evaluated on the common test data of 90 sentences. The models were trained on the data from each subsequent annotation round and the training data were incremented after each training step.

The evaluation process was repeated three times and subsequently mean values were computed. The best F-score value was achieved using the complete training data set. The F-score for detecting two categories of pathological phenomena, i.e. *short* and *long*, is 0.37 (with a variance of 0.07). The F-score for detecting only *short* pathological phenomena is higher, with a value of 0.55 (with a variance of 0.05). Obviously the performance variance is higher for the detection of both categories together than for the detection of the *short* category only.

The results we achieved for the seed set and for the complete 13 annotation rounds are presented in Table 2. The numbers reveal that we could extract both categories with a performance of 0.34 recall, 0.41 precision and 0.37 F-score (before the start of the AL rounds the performance was 0.27 recall, 0.36 precision and 0.31 F-score). Obviously, the F-score value after the completion of the 13 AL rounds is higher than the F-score value at the start of the AL rounds. However, the increase by six percentage points and the overall performance value of 0.37 F-score are modest. This might be due to the intrinsic complexity of the *Pathological Phenomena* category *long*. When we evaluated the *short* category only, after the completion of 13 AL rounds we achieved a competitive performance of 0.48 recall, 0.66 precision and 0.55 F-score. Figures 1 and 2 visualize the learning progress (blue line) and illustrate that the increase in F-score performance can be fitted by a logarithmic function using the least squares method (red line).

Table 2: The recall, precision and F-score values represent mean values resulting from a 3-trials evaluation on test data. The evaluation was performed entity-wise and required the exact matching of entity spans.

		Seed Set			13 AL rounds		
Pathological Phenomena	Test size (sentences)	Recall	Precision	F-score	Recall	Precision	F-score
short & *long*	90	0.27	0.36	0.31	0.34	0.41	**0.37**
short	90	0.32	0.54	0.40	0.48	0.66	**0.55**

6 Discussion

In this paper, we presented an annotation project targeting *Pathological Phenomena* that was carried out using an Active Learning-based annotation approach. AL was a strategic choice based on previously gathered empirical evidence indicating that AL-based annotation yields (often huge) benefits in terms of annotation efficiency without degrading annotation effectiveness, compared with a baseline approach, usually random sampling (see Section 2). Since we do not know of any published work that cites data in favor of random sampling in direct comparison with AL, we refrained from coding of our own random sampling baseline (by means of which we could have exactly quantified the benefits AL brings in our annotation context).

The F-score values for training the NER tagger for *Pathological Phenomena* are reasonably low. This holds for the annotation of the *short* mention category (0.55) which typically relates to diseases but even more so for the combination of *short* and *long* mentions (0.37). Three explanations might be worthwhile to put these results in perspective. First, we used an off-the-shelf NER tagger in our pipeline (which is much more sensitive to gene/protein recognition and normalization, see Wermter et al. [23]) and applied it on an *as is* basis, while the adaption to the medical domain is pending. Second, the number of tokens we are dealing with (on the order of 13,000) is still very low, given evidence

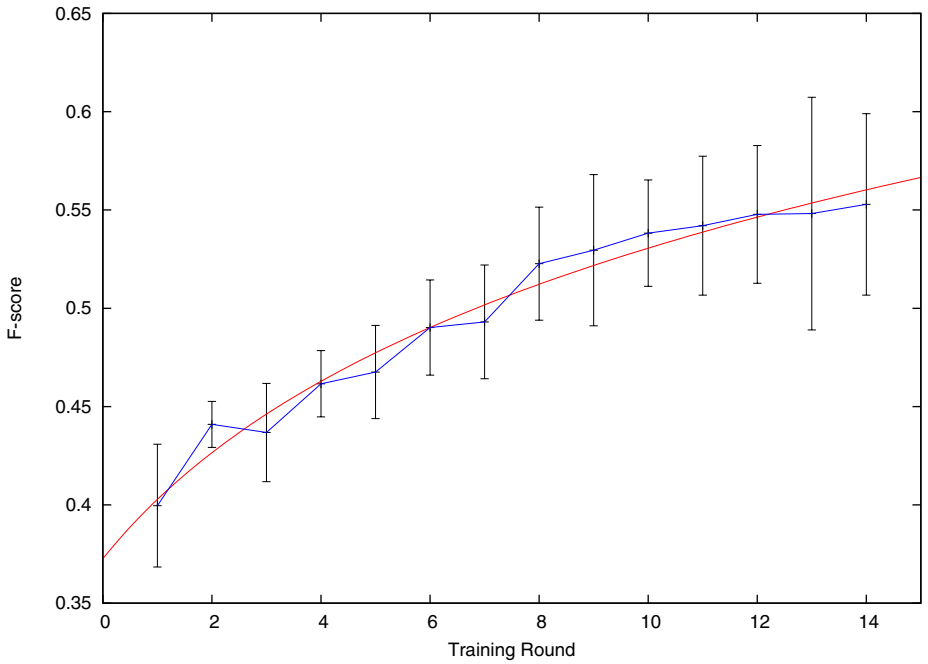

Figure 1: Learning progress for *short* pathological phenomena (blue line) and a logarithmic function fitted to the data using the least squares method (red line).

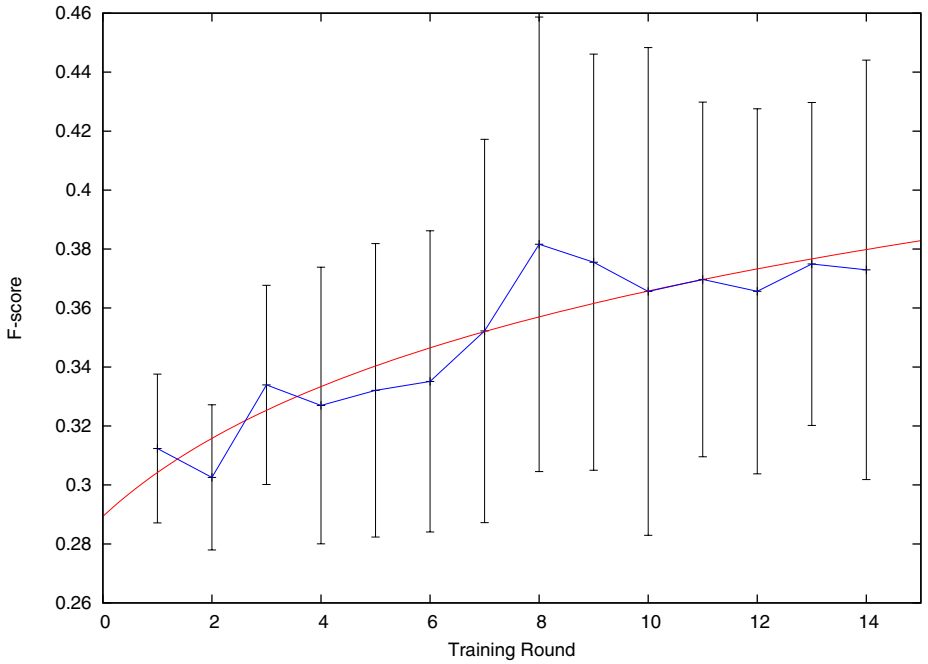

Figure 2: Learning progress for *short* and *long* pathological phenomena (blue line) and a logarithmic function fitted to the data using the least squares method (red line).

that near-optimal results usually require more than double the amount of tokens, for the newspaper domain as well as the biomedical domain, at least.[10] Finally, the *long* category is still hard to tackle, since it ranges from (complex) noun phrases to full sentences. This also becomes evident from the enormous variances we encounter for the combined *long* and *short* annotations at almost all measurement points, i.e. training rounds (see Figure 2). These fuzzy data make it hard to reliably estimate the slope of the learning curve. We soon hope to generate further evidence that the dual category might pay off, on a larger corpus scale, at least, using an adapted version of our NER tagger on a significantly enhanced corpus.

7 Conclusions

In this paper, we make two contributions to the field of corpus annotation in medicine. First, two new units of annotation are proposed — short *versus* long mention spans. They are supposed to capture the intrinsic complexity of the coding of *Pathological Phenomena*, an entity type with quite sloppy linguistic appearance and underlying semantics. Short mention spans primarily target terminologically well-defined diseases, whereas long mention spans focus on descriptive statements indicating (quasi-)pathological phenomena and, very broadly, norm-deviant states and observations. The latter class of utterances is rarely dealt in medical document annotations despite the fact that it is highly relevant for any sort of diagnostic and therapeutic task. Second, a novel annotation strategy was chosen, Active Learning, which unlike random sampling introduces a sampling bias by focusing on the most uncertain annotation exemplars during the selection phase for items to be annotated. Based on these decisions, we generated PATHOJEN, a text corpus currently composed of 900 sentences (13,000 tokens) from MEDLINE abstracts, with roughly 1,300 *short* and 700 *long* mention span annotations.

We started the development of the PATHOJEN corpus on the premise that Active Learning-based annotation would yield benefits similar to the ones we had previously found for standard named entity categories in the newspaper domain (*Persons*, *Locations*, etc.) and the biological domain (*Genes*, *Proteins*, etc.).[10] However, the currently available learning curves for annotating pathological phenomena in the PATHOJEN corpus do not seem to match this expectation. The steep increase we had determined for the *Persons* or *Locations* cases is in stark contrast with the moderate logarithmic growth of the learning curves for *Pathological Phenomena*. As a caveat, we point to the limited size of the token set under scrutiny and the immense ranges of variances we determined for our measurements. Since AL usually pays off only at volumes more than double the amount of tokens we have considered up until now, in order to strengthen our observations on the slope of the learning curves we have to increase the size of our annotated corpus.

There are further speculations to explain the differences in the shapes of the curves. First, the learning task might indeed be intrinsically harder for *Pathological Phenomena* than for the *Persons* or *Locations* cases. This might already be true for *short* encodings of *Pathological Phenomena* but is certainly true for *long* ones that are extremely heterogeneous in their linguistic appearance. Second, for measuring the learning curves in our AL experiments we used a named entity tagger, without any adaptions, that was originally optimized for biological named entity recognition (*Genes* and *Proteins*, in particular),[23] and not for the recognition of medical named entities such as *Pathological Phenomena*. Hence, tuning the features of this classifier might yield boosts in performance that are independent from the presumed complexity of the recognition task for *Pathological Phenomena*.

Altogether, our experiments indicate that, first, the very optimist promises of AL seem to be under stress in the area of *Pathological Phenomena*. Second, the sloppy nature of pathological phenomena, as they occur in verbal statements, seems to constitute an area of inherent complexity that is a true challenge for any annotation effort. Future work might tell us which of the two lines of thought (or both) are strong impediments to the generation of more adequately semantics-annotated clinical corpora.

Acknowledgments. This work is 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).

References

1. Hahn Udo, Beisswanger Elena, Buyko Ekaterina, Faessler Erik, Traumüller Jenny, Schröder Susann et al. Iterative refinement and quality checking of annotation guidelines: how to deal effectively with semantically sloppy named entity types, such as pathological phenomena. In *LREC 2012 – Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3881–3885. May 21-27, 2012, Istanbul, Turkey, 2012.

2. Settles Burr. Active learning literature survey. Technical Report 1648, University of Wisconsin - Madison, Computer Sciences Technical Report 1648, 2010. Computer Sciences Technical Report 1648.

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

4. Roberts Angus, Gaizauskas Robert, Hepple Mark, Davis Neil, Demetriou George, Guo Yikun et al. 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. 10-14 November 2007, Chicago, IL, USA, 2007.

5. Uzuner Özlem, South Brett R., Shen Shuying and DuVall Scott L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

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

7. Leaman Robert, Gonzalez Graciela and Miller Christopher. 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. November 8-10, 2009, Seogwipo-si, Jeju Island, South Korea, 2009.

8. Ogren Philip V., Savova Guergana K. and Chute Christopher G. Constructing evaluation corpora for automated clinical named entity recognition. In *LREC 2008 – Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 3143–3150. 29-30 May 2008, Marrakech, Morocco, 2008.

9. Pestian John P., Brew Christopher, Matykiewicz Pawel, Hovermale D. J., Johnson Neil, Cohen K. Bretonnel et al. 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. June 29, 2007, Prague, Czech Republic, 2007.

10. Tomanek Katrin, Wermter Joachim and Hahn Udo. 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 486–495. June 28-30, 2007, Prague, Czech Republic, 2007.

11. Reichart Roi, Tomanek Katrin, Hahn Udo and Rappoport Ari. Multi-task active learning for linguistic annotations. In *ACL 2008 – Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 861–869. June 15-20, 2008, Columbus, OH, USA, 2008.

12. Tomanek Katrin and Hahn Udo. Reducing class imbalance during active learning for named entity annotation. In *K-CAP 09 – Proceedings of the 5th International Conference on Knowledge Capture*, pages 105–112. September 1-4, 2009, Redondo Beach, CA, USA, 2009.

13. Tomanek Katrin, Hahn Udo, Lohman Steffen and Ziegler Jürgen. A cognitive cost model of annotations based on eye-tracking data. In *ACL 2010 – Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167. 11-16 July 2010, Uppsala, Sweden, 2010.

14. Liu Y. Active learning with Support Vector Machine applied to gene expression data for cancer classification. *Journal of Chemical Information and Modeling*, 44(6):1936–1941, 2004.

15. Kim Seokhwan, Song Yu, Kim Kyungduk, Cha Jeong-Won and Lee Gary Geunbae. MMR-based active machine learning for bio named entity recognition. In *HLT-NAACL 2006 – Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, Companion Volume: Short Papers*, pages 69–72. June 4-9, 2006, New York, NY, USA, 2006.

16. Tsuruoka Yoshimasa, Tsujii Jun'ichi and Ananiadou Sophia. Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC Bioinformatics*, 9(Suppl 11):S8, 2008.

17. Doyle Scott, Monaco James, Feldman Michael, Tomaszewski John and Madabhushi Anant. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC Bioinformatics*, 12:424, 2011.

18. Chen Yukun, Mani Subramani and Xu Hua. Applying active learning to assertion classification of concepts in clinical text. *Journal of Biomedical Informatics*, 45(2):265–272, 2012.

19. Mohamed Thahir P., Carbonell Jaime G. and Ganapathiraju Madhavi K. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, 11(Suppl 11):S57, 2010.

20. Olsson Fredrik and Tomanek Katrin. An intrinsic stopping criterion for committee-based active learning. In *CoNLL 2009 – Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 138–146. June 4-5, 2009, Boulder, CO, USA, 2009.

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

22. Hahn Udo, Buyko Ekaterina, Landefeld Rico, Mühlhausen Matthias, Poprat Michael, Tomanek Katrin et al. An overview of JCoRe, the JULIE Lab UIMA component repository. In *Proceedings of the LREC 2008 Workshop 'Towards Enhanced Interoperability for Large HLT Systems - UIMA for NLP'*, pages 1–7. 31 May 2008, Marrakech, Morocco, 2008.

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