Open-domain Anatomical Entity Mention Detection

Tomoko Ohta [1] **Sampo Pyysalo** [1] **Jun'ichi Tsujii** [2] **Sophia Ananiadou** [1]
[1]National Centre for Text Mining and University of Manchester,
Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, UK
[2]Microsoft Research Asia, Beijing, China
okap.tiffany@gmail.com, sampo.pyysalo@gmail.com
jtsujii@microsoft.com, sophia.ananiadou@manchester.ac.uk

Abstract

Anatomical entities such as *kidney*, *muscle* and *blood* are central to much of biomedical scientific discourse, and the detection of mentions of anatomical entities is thus necessary for the automatic analysis of the structure of domain texts. Although a number of resources and methods addressing aspects of the task have been introduced, there have so far been no annotated corpora for training and evaluating systems for broad-coverage, open-domain anatomical entity mention detection. We introduce the AnEM corpus, a domain- and species-independent resource manually annotated for anatomical entity mentions using a fine-grained classification system. The corpus texts are selected randomly from citation abstracts and full-text papers with the aim of making the corpus representative of the entire available biomedical scientific literature. We demonstrate the use of the corpus through an evaluation of the broad-coverage MetaMap tagger and a CRF-based system trained on the corpus data, considering also a combination of these two methods. The combined system demonstrates a promising level of performance, approaching 80% F-score for mention detection for a relaxed matching criterion. The corpus and other introduced resources are available under open licences from http://www.nactem.ac.uk/anatomy/.

1 Introduction

Entity mention detection is a prerequisite for most efforts to systematically analyse and represent the structure of scientific discourse. In the life sciences, a comprehensive analysis must include entities at multiple levels of biological organization, from the molecular to the organism level. The detection of references to *anatomical entities* such as "*kidney*" and "*blood*" is thus required for the automatic structured analysis of biomedical scientific text.

Although a wealth of lexical and ontological resources covering anatomical entities are available (Rosse and Mejino, 2003; Smith et al., 2007; Bodenreider, 2004; Haendel et al., 2009), such resources do not alone confer the ability to reliably detect mentions of anatomical entities in natural language (Gerner et al., 2010a; Travillian et al., 2011; Pyysalo et al., 2012b). To support the development and evaluation of reliable anatomical entity mention detection methods, corpus resources annotated specifically for the task are necessary.

In this study, we aim to create a reference standard for evaluating methods for anatomical entity mention detection and for training machine learning-based methods for the task. We seek to select a set of texts that are representative of the relevant scientific literature, i.e. *open-domain* in the sense of avoiding bias toward, for example, specific species, levels of biological organization (e.g. subcellular or gross anatomy), parts of documents (e.g. abstracts), or subdomains of life science. In support of our annotation, we draw on a granularity-based, species-independent upper-level ontology of anatomy as well as relevant species-specific ontological resources.

The overall aim of our efforts is to create methods and resources for comprehensive event-based analysis (Ananiadou et al., 2010) of biomedical scientific discourse involving anatomy-level entities and processes. In aiming to establish a stable basis for anatomical entity mention detection, the present study is an important step toward this goal.

		Label	Ontology classes	Examples
Anatomical entity	Anatomical structure	ORGANISM SUBDIVISION	organism subdivision CARO	*head, limb*
		ANATOMICAL SYSTEM	anatomical system CARO	*vascular system*
		ORGAN	compound organ CARO	*liver, heart*
		MULTI-TISSUE STRUCTURE	multi-tissue structure CARO	*artery*
		TISSUE	portion of tissue CARO	*epithelium*
		CELL	cell CARO	*epithelial cell*
		DEVELOPING ANATOMICAL STRUCTURE	developing anatomical structure UBERON	*embryo*
		CELLULAR COMPONENT	cellular component GO	*mitochondrion*
		ORGANISM SUBSTANCE	portion of organism substance CARO	*blood*
		IMMATERIAL ANATOMICAL ENTITY	immaterial anatomical entity CARO	*lumen*
		PATHOLOGICAL FORMATION	–	*carcinoma*

Table 1: Annotations targets with applied label, corresponding ontology classes, and common examples.

2 Corpus Annotation

2.1 Ontological Basis

Following our previous efforts on anatomical entity classification (Pyysalo et al., 2012b), we base our definition of annotated mention scope, the subdivision of anatomical entities into classes, and the class labels applied in our annotation primarily on the Common Anatomy Reference Ontology (CARO) (Haendel et al., 2008). CARO is a small, species-independent ontology of anatomical entities based on the upper-level structure of the Foundational Model of Anatomy (FMA) ontology of human anatomy (Rosse and Mejino, 2003; Rosse and Mejino, 2008). CARO has been proposed as a standard for unifying the upper-level structure of the various existing species-specific ontologies and is adopted by many of the over 40 ontologies involving the anatomy domain in the Open Biomedical Ontologies (OBO) foundry[1] (Smith et al., 2007). CARO adheres to disjoint classes and single inheritance, and divides anatomical structures primarily by granularity (Kumar et al., 2004), a systematic notion familiar to those working in the life sciences.

Although we draw primarily on CARO, we follow the well-established cellular component subontology of the Gene Ontology (GO) (Ashburner et al., 2000) in grouping sub-cellular structures under a single upper-level category. For developing structures that resist granularity-based categorization due to occupying different levels at different stages of development, we adopt a separate DEVELOPING ANATOMICAL STRUCTURE category, as done also in e.g. Uberon (Haendel et al., 2009).

[1] http://obofoundry.org/

2.2 Annotation Scope

We diverge from the scope of anatomy ontologies in two important aspects in our annotation.

First, ontologies of anatomy commonly incorporate everything from molecules to whole organisms within their scope. However, in entity mention detection, many molecular level anatomical entities fall within the scope of the established gene/protein mention detection tasks (e.g. (Kim et al., 2004; Tanabe et al., 2005)), and whole organism mentions similarly largely within what is covered by existing methods and resources for organism mention detection (Gerner et al., 2010b; Naderi et al., 2011). To avoid overlap with established tasks and to focus on the novel aspects of anatomical entity mention detection, we exclude biological macromolecules and mentions of organism names from the scope of our annotation, as argued in (Pyysalo et al., 2012b).

Second, these ontologies typically represent *canonical* anatomy, an idealized state that is rarely (if ever) encountered in reality (Bada and Hunter, 2011). As our annotation is intended to cover references to real-world anatomy, we explicitly include in the scope of our annotation also healthy as well as pathological variants of canonical anatomy. We include also entities derived from these anatomical entities through (planned) processing such as surgical or laboratory procedures, even when these processed entities are no longer properly part of the original organism. Finally, we annotate pathological formations such as scars and carcinomas that are part of individual organisms but have no correspondence in canonical anatomy (Smith et al., 2005).

Table 1 presents the class labels applied in the annotation with the corresponding ontology classes.

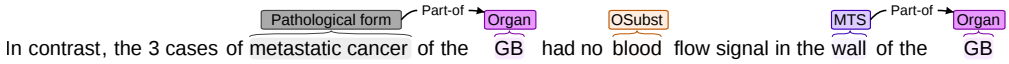

Figure 1: Example sentence with annotation. OSUBST and MTS abbreviate for ORGANISM SUBSTANCE and MULTI-TISSUE STRUCTURE, respectively.

2.3 Representation

The primary corpus annotation marks mentions of anatomical entities as contiguous spans of characters in text, each of which is assigned a type (Figure 1). As the CARO-based categorization has comprehensive coverage and disjoint classes, each annotation can be assigned exactly one type (class label).

In addition to identifying and typing anatomical entity mentions, we further apply binary attributes ("flags") marking the following characteristics of each mention:

DEVELOPING developing variant of anatomical entity, e.g. *fetal liver*

PATHOLOGICAL pathological variant of anatomical entity, e.g. *carcinoma cell*

PLANT anatomical entity that is part of a plant (member of the *Viridiplantae* kingdom), e.g. *roots, leaf*

PROCESSED variant of anatomical entity that has undergone planned processing, e.g. *tissue specimen*

Any combination of attributes can apply to a single mention. These attributes allow the identification of subsets of annotations that may be out of scope for some efforts (e.g. pathological or processed entities) and facilitate the analysis of mention detection system performance by identifying particular problematic categories.

2.4 Annotation Criteria

In very brief summary, we annotate spans of text that refer to anatomical entities as defined above. Mentions that involve only metaphorical senses of such entities ("*on the other* hand") or artificial analogues ("*artificial* heart") are not annotated.

The primary targets of our annotation are anatomical entity names (e.g. "*lymphocyte*") and nominal mentions of anatomical entities (e.g. "*muscle tissue*"). Both names and nominal mentions are annotated similarly, without distinction. We exclude pronouns (*it*, *that*) from annotation even when they unambiguously refer to an anatomical entity; we consider the identification and resolution of such mentions part of the distinct coreference resolution task (see e.g. Pradhan et al. (2011)).

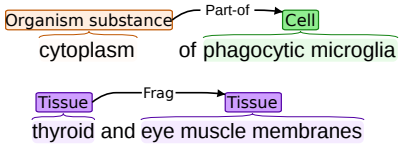

Figure 2: Part-of relation marking entity mention spanning a prepositional phrase (above) and Frag relation marking coordination with ellipsis (below).

In addition to names and nominal mentions, we mark adjectives that have an unambiguous sense of relating to a specific anatomical entity. Thus, for example, both "*kidney*" and "*renal*" (relating to the kidneys) are annotated as ORGAN in expressions such as "*kidney failure*" and "*renal failure*". The choice to annotate adjectival references is motivated by the expected needs of applications making use of automatically detected anatomical entity mentions. For example, for semantic search targeting documents relating to organ failure, a document discussing "*renal failure*" is obviously relevant and should be recovered.

Syntactically, annotations mainly cover base noun phrases without determiners, i.e. nouns with premodifiers relevant to identifying the specific anatomical entity referred to. We exclude noun phrase postmodifiers such as prepositional phrases from the span of single annotations, but apply a separate level of annotation for *part-of* relations that allow such alternate spans to be recovered when they identify an anatomical entity (Figure 2 top). Similarly, we decompose coordinated references to anatomical entities involving ellipsis to non-overlapping spans, but mark the cases using a *frag*(*ment*) relation type (Figure 2 bottom). (Due to space considerations, we omit detailed discussion of these relation annotations.) Together with the properties described in Section 2.3, these constraints assure that any single token is assigned at most one class label and allow the annotation to be repre-

Task	Matching criterion		
	Strict	Left boundary	Right boundary
Mention detection (single class)	89.2%/ 82.0%/ 85.4%	93.0%/ 85.5%/ 89.1%	94.6%/ 86.9%/ 90.6%
Detection and classification (multi-class)	85.6%/ 78.7%/ 82.0%	87.0%/ 80.0%/ 83.3%	90.2%/ 82.9%/ 86.4%

Table 2: Inter-annotator agreement results (precision / recall / F-score).

sented in the standard BIO format and to be straightforwardly applied with many existing entity mention taggers.

By contrast to previously introduced domain resources for e.g. molecular entity and organism mention detection (Tanabe et al., 2005; Gerner et al., 2010b), we do not incorporate any specificity constraints in our annotation criteria. That is, non-specific expressions such as "*tissue*" and "*organ*" are marked identically to specific ones such as "*epithelium*" and "*heart*". This choice seeks to assure the generality of the task and methods for addressing it.

2.5 Text Selection

Texts for the corpus were drawn from two sources: the PubMed[2] database of publication abstracts, and the PubMed Central[3] (PMC) Open Access subset of full-text publications. PubMed, containing more than 20 million citations, has a very broad coverage of domain scientific texts but is limited to publication abstracts, while PMC has lower coverage but does provide over 400,000 full-text documents under open licenses. By sampling both sources, we seek to assure the corpus is relevant to IE efforts regardless of their choice of texts.

To avoid bias toward e.g. subdomains of biology or specific species, we selected texts from both sources by random sampling. For PubMed, we simply selected a random set of citations and extracted their abstract and title texts. For PMC, we initially extracted all non-overlapping section texts (PMC XML <sec> elements) as well as caption texts (<caption> elements), and then selected a random set of extracts. This selection seeks to maximize the diversity of the texts in the full-text section of the corpus, and the selection of extracts larger than isolated sentences aims to allow the corpus to be used to study methods making use of broader context, e.g. by incorporating constraints such as one sense per discourse (Gale et al., 1992).

We selected a total of 500 documents using this protocol, half from PubMed and half from PMC document extracts. (Descriptive statistics of the *abstracts* and *full-text extracts* subcorpora are given later in Table 3.)

2.6 Annotation Process

Primary annotation was created by a PhD biologist with extensive experience in domain information extraction and text annotation (TO). The use of any relevant resources, such as the full article being annotated or species-specific anatomy ontologies in the OBO foundry, was encouraged for resolving unclear or ambiguous cases during annotation. Initial annotation was produced entirely manually. To further assure the quality of the annotation, a series of automatic tests was performed and used as the basis of a further manual round of revision.[4] Annotation guidelines were initially created based on those created by our previous domain-specific effort (Pyysalo et al., 2012a) and revised throughout the annotation effort to document specific decisions made during annotation. The annotations were created using the BRAT annotation tool (Stenetorp et al., 2012).

To evaluate the annotation consistency, we performed an inter-annotator agreement (IAA) experiment. After brief training with annotation guidelines provided by the primary annotator, a random 10% of the corpus was independently annotated by a PhD computer scientist with experience in domain text annotation and anatomy ontologies (SP). IAA was evaluated using the same criteria as applied in experiments (see Section 3.4), holding the primary annotation as gold. The results are shown in Table 2. We find very good agreement both for mention detection (ignoring classification) as well as for the full task, indicating that the task is well defined and the annotation consistency high.

[2] http://pubmed.com

[3] http://www.ncbi.nlm.nih.gov/pmc/

[4] No automatically suggested annotations were incorporated into the corpus without manual verification.

3 Methods

We next present the methods applied in our anatomical entity mention detection experiments. We aim to evaluate the capacity of the newly annotated corpus to support reliable mention detection and to establish initial baseline results for the newly introduced resource, and thus focus only on relatively straightforward applications of existing methods.

3.1 MetaMap

MetaMap[5] (Aronson, 2001) is a tool capable of detecting mentions of concepts from the extensive UMLS Metathesaurus (Bodenreider, 2004) in text. The metathesaurus and MetaMap have broad coverage of concepts relevant to biology and medicine and provide a categorization of concepts into 133 semantic types, ranging from `Amino Acid` to `Health Care Activity` to `Vertebrate`, many directly relevant to anatomical entities. MetaMap is a key component of the process used by the National Library of Medicine (NLM) to index publications in the PubMed database and has been applied in numerous other information extraction and information retrieval tasks (Aronson and Lang, 2010).

In initial experiments, we applied MetaMap to training set documents to identify the subset of the 133 semantic classes relevant to anatomy, selecting 14 classes (including e.g. `Cell`, `Tissue` and `Body Substance`) for final experiments.[6] During testing, we used command-line arguments to restrict output to the selected semantic classes. The core tagging functionality of MetaMap is rule-based, and it does not support training on tagged data for concept mention detection. With the exception of the semantic class selection, the evaluation of MetaMap reflects an "off-the-shelf" application of the general-purpose tool.

3.2 CRF tagging

Conditional Random Fields (CRF) (Lafferty et al., 2001) are graphical models that are frequently applied to sequence labeling tasks, and CRFs form the basis of state-of-the-art methods for many entity mention tagging tasks. We performed experiments using the NERsuite entity mention recognition toolkit, based on the CRFsuite implementation of CRFs (Okazaki, 2007). NERsuite provides an extensive set of features applied in entity mention detection, allowing the tool to achieve performance competitive with state-of-the-art methods for many biomedical domain tasks through retraining without task-specific adaptation[7]. Retraining the tool for new tasks is also straightforward, allowing application to new tasks with modest effort.

We set the L_2 regularization parameter of the learning method using held-out evaluation with training set data, picking out of a set of values 2^n ($n \in \mathbb{Z}$) the one giving best performance.[8] Other learning method parameters were left at default values.

3.3 System combination

As a third system, we apply a straightforward combination of the MetaMap and CRF tagging systems, where we initially tag the data using MetaMap and then incorporate the classes assigned by MetaMap as features for training and testing with NERsuite (stacking). More specifically, we create a BIO-tagged version of MetaMap output segmented to match NERsuite tokenization, and assign each token the BIO tag based on the MetaMap semantic type code (e.g. `B-cell`) as a feature.

Excepting for the addition of these MetaMap-derived features, NERsuite is applied as described above (Section 3.2).

3.4 Experimental setting

We split the corpus data into two primary parts: a training set consisting of 60% of the documents and a test set of the remaining 40%. The data splits were performed independently for the two subcorpora (abstracts and full-text extracts), using stratified sampling to assure broadly comparable statistical properties between the sets. The test set was held out during development and only applied for the final experiments.

[5] `http://metamap.nlm.nih.gov/`

[6] In brief, we tagged the training data with MetaMap, extracted the subset of semantic classes giving more than 5% precision against the gold annotations, and manually analysed these to select this subset. The selected classes are detailed in supplementary material available on the project webpage.

[7] `http://nersuite.nlplab.org/`

[8] Specifically, $C_2 = 2^{-5}$ was selected.

Source	Item	Dataset Train	Test	Total
Abst.	Document	150	100	250
	Word	28,960	18,199	47,159
	Entity	1,182	764	1,946
FTE	Document	150	100	250
	Word	26,306	17,955	44,261
	Entity	697	492	1,189
Total	Document	300	200	500
	Word	55,266	36,154	91,420
	Entity	1,879	1,256	3,135

Table 3: Overall corpus statistics. Statistics given separately for the abstracts (abst.) and full-text extracts (FTE) subcorpora as well as for the total.

Type	Count
CELL	776
MULTI-TISSUE STRUCTURE	639
ORGAN	381
PATHOLOGICAL FORMATION	368
ORGANISM SUBSTANCE	291
CELLULAR COMPONENT	199
TISSUE	169
ORGANISM SUBDIVISION	162
IMMATERIAL ANATOMICAL ENTITY	60
ANATOMICAL SYSTEM	51
DEVELOPING ANATOMICAL STRUCTURE	39

Table 4: Annotation statistics by type.

We perform experiments in two settings: a single-class setting where the task is restricted to the detection of anatomical entity mentions without classification, and a multi-class setting where the correct class label must further be assigned to each detected mention. As MetaMap uses UMLS semantic classes that do not fully align with the applied CARO-based classes, MetaMap is only applied in the single-class setting.

For evaluation, we adopted the protocol, criteria and metrics of the established BioNLP/JNLPBA shared task 2004 (Kim et al., 2004). To assure compatibility, we created our evaluation tool on the basis of the shared task evaluation script. The evaluation is thus based on entity-wise (microaverage) precision/recall/F-score metrics, and tagging performance is separately evaluated under *strict match*, *left boundary match* and *right boundary match* criteria. In the former setting, a predicted entity must exactly match the extent of a gold standard entity, while in the latter two settings, it is enough that the left/right boundary matches.

3.5 Format

The annotation is distributed in the standard column-based BIO format applied for e.g. CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) and JNLPBA (Kim et al., 2004) data, among other established datasets.

4 Results

4.1 Corpus statistics

Table 3 presents the overall corpus statistics. We note that the abstracts and full-text extracts (FTE) subcorpora are of comparable size in terms of their word counts, but the number of annotations is 1.6 times higher in the abstracts subcorpus (1.5 correcting for number of words). This difference in anatomical entity mention density between abstracts and full texts parallels the findings of Cohen et al. (2010) on the relative density of gene, drug and disease mentions. We further note that the estimated density of anatomical entity mentions in abstracts (approx. 41 per 1000 words) and full texts (27 per 1000) are broadly comparable to the gene mention density estimates of Cohen et al. (61 and 47 for abstracts and full texts, respectively).

Table 4 presents a breakdown by annotation type. There are large differences in the number of annotations by type, with the majority class CELL outnumbering the rarest type 20-fold. While the total number of annotated examples is likely to be sufficient for training machine learning-based taggers and most of the classes contain a respectable number of examples, the statistics suggest that the least frequently annotated types may represent challenges for learning.

4.2 Entity Mention Detection

Table 5 presents the experimental results for anatomical entity mention detection (single-class). In terms of F-score, we find the same ranking of the three methods for all three criteria, with the CRF-based tagger outperforming the rule-based MetaMap, and the combination method outperforming its components. Although it is not surprising that a dedicated machine learning-based system is capable of outperforming a general-purpose, largely rule-based system, this result does reflect positively on both the

Method	Matching criterion		
	Strict	Left boundary	Right boundary
MetaMap	50.78% / 64.49% / 56.82%	54.67% / 69.43% / 61.17%	58.18% / 73.89% / 65.10%
NERsuite	77.98% / 52.15% / 62.50%	81.43% / 54.46% / 65.27%	90.00% / 60.19% / 72.14%
MetaMap + NERsuite	82.09% / 62.42% / 70.92%	84.61% / 64.33% / 73.09%	90.68% / 68.95% / 78.34%

Table 5: Overall single-class anatomical entity mention detection results (precision / recall / F-score).

Method	Matching criterion		
	Strict	Left boundary	Right boundary
NERsuite	72.07% / 42.12% / 53.17%	72.75% / 42.52% / 53.67%	85.69% / 50.08% / 63.22%
MetaMap + NERsuite	75.41% / 51.75% / 61.38%	76.45% / 52.47% / 62.23%	83.99% / 57.64% / 68.37%

Table 6: Overall anatomical entity mention detection and classification results (precision / recall / F-score).

consistency of the annotation as well as the sufficiency of the size of the newly introduced corpus. In this application, we find that MetaMap tends to favor recall over precision – perhaps reflecting its focus on IR applications (Aronson and Lang, 2010) – while the trained machine learning-based models are clearly biased in favor of high precision.

As expected on the basis of the results of previous evaluations using similar experimental setups (Kim et al., 2004), results are notably better under the relaxed matching criteria. In particular, requiring only the right boundaries of annotations to match yields F-scores nearly 10% points higher than under strict matching. Recalling that the annotations primarily mark base noun phrases, this suggests that the systems comparatively frequently identify the head word of an anatomical entity mention correctly but differ from gold annotation regarding the choice of premodifiers included in the span of the annotation. As limited variation in premodifier selection is arguably acceptable for many applications and relaxed matching criteria are frequently applied in domain tagging tasks (Kim et al., 2004; Wilbur et al., 2007), we propose to consider performance under the relaxed *right boundary match* criterion as the primary result for evaluation using the new corpus.

Table 6 presents the results for anatomical entity mention detection and classification using the 11-class categorization used in annotation.[9] While performance in terms of F-score is approximately 10% points lower than for the single-class task, this drop is comparatively modest given the large number of distinct classes, indicating that the number of annotations of most individual classes is sufficient for learning.

While these initial results are not as high as for established entity mention detection tasks in the domain (Wilbur et al., 2007; Rebholz-Schuhmann et al., 2011), we consider the level of performance quite good given the many new challenges relating to the task. Further, as the mention detection methods were also applied with only modest specific adaptation to the task, we believe there remain many opportunities for further development of methods for the task.

4.3 Discussion

Many commonly targeted mention types in both the "general" and the biological domain are frequently characterized by obvious surface features: the names of people and locations are capitalized in many languages, as are genera in scientific species' names, and many gene and chemical names have comparable features distinguishing them from common nouns (consider e.g. *p53*, *IgE*, *c-myc*, *Ca2+*, *H2SO4*). By contrast, many typical anatomical entity mentions are common noun compounds lacking obvious distinguishing surface features. This fact likely contributes to the comparatively low performance of the CRF-based tagger when applied without support from lexical resources.

A further challenge that arises comparatively frequently in anatomical entity mention detection is ambiguity between entity mentions and other words sharing the same surface form. For example, while *Barack Obama*, *Sweden*, *p53* and *H2SO4* can be

[9] Note that evaluation using MetaMap only is not possible as its semantic classes differ from those used in the annotation.

safely identified as mentions of a person, country, gene, and chemical without reference to context, *face* should not be marked as an anatomical entity mention in *face the facts*, nor should *Airways* in *British Airways*. Thus, approaches relying on simple matching against lexical resources will not suffice for accurate anatomical entity mention detection.

Our evaluation results demonstrated a clear advantage to combining detection based on lexical resources with machine learning-based tagging, an approach we believe will be key to the further development of reliable anatomical entity mention tagging that we will seek to explore in detail in future work. To facilitate analysis of the performance of the methods, we provide the predictions of each method in supplementary data on the project homepage.

5 Related work

A number of domain corpora such as GENIA (Ohta et al., 2002), BioInfer (Pyysalo et al., 2007), and the recently introduced CellFinder corpus (Neves et al., 2012) include annotation for at least some classes of anatomical entities. However, such corpora typically cover only specific subdomains of the literature, such as transcription factors in human blood cells (GENIA), protein-protein interactions (BioInfer), or stem cells (CellFinder). To the best of our knowledge, this is the first effort introducing a corpus annotated for anatomical entity mentions that specifically aims to be representative of the entire available literature. We note that there is a well-established precedent to this goal: sentences for the *de facto* standard corpus for gene/protein name recognition, GENETAG (Tanabe et al., 2005), were similarly selected from PubMed abstracts without domain restrictions.

The BioNLP/JNLPBA shared task 2004 (Kim et al., 2004) targeted the detection of mentions of five types of biological entities, including two that would fall within in the scope of our CELL annotation ("Cell type" and "Cell line"). Other than this comparatively early shared task, collaborative domain efforts such as BioCreative (Krallinger et al., 2008) and CALBC (Rebholz-Schuhmann et al., 2011) have not targeted anatomical entity mentions.

Some recent studies have considered the use of ontological resources for the detection of anatomical entity mentions in natural language expressions. In previous work (Pyysalo et al., 2012b), we studied the classification of isolated noun phrases extracted from PubMed to identify anatomy terms. Travillian et al. (2011) considered two lexical matching applications to detect anatomical entities from two OBO resources in user-provided terms. However, these efforts have not involved the annotation or detection of mentions in context, which we view as critical for real-world entity mention detection method development and evaluation.

6 Conclusions

We have introduced a manually annotated corpus for open-domain anatomical entity mention detection, consisting of 500 documents (over 90,000 words) drawn from publication abstracts and full texts. The primary corpus annotation consists of the identification of over 3,000 references to both healthy and pathological anatomical entities, marked using a detailed 11-class categorization based on established biomedical domain ontologies. We demonstrated the use of the new corpus through a comparative evaluation of MetaMap, a general semantic class tagger; NERsuite, a CRF-based machine learning system; and a stacked combination of the two, finding that under a relaxed matching criterion, the combination approaches 80% F-score at mention detection and 70% F-score at mention detection and classification. This level of performance is encouraging for a first application and suggests that reliable open-domain anatomical entity mention detection is not an unrealistic target.

We hope that the introduced corpus can serve as a reference standard for the further development and evaluation of methods for anatomical entity mention detection. This corpus, the introduced evaluation tools, and other resources created in this study are made available under open licences from `http://www.nactem.ac.uk/anatomy/`.

Acknowledgments

This work was funded by UK Biotechnology and Biological Sciences Research Council (BBSRC) under project Automated Biological Event Extraction from the Literature for Drug Discovery (reference number: BB/G013160/1).

References

S. Ananiadou, S. Pyysalo, J. Tsujii, and D.B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

A.R. Aronson and F.M. Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

A.R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of AMIA*, pages 17–21.

M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25:25–29.

M. Bada and L. Hunter. 2011. Desiderata for ontologies to be used in semantic annotation of biomedical documents. *Journal of Biomedical Informatics*, 44(1):94–101.

O. Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

K.B. Cohen, H. Johnson, K. Verspoor, C. Roeder, and L. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics*, 11(1):492.

W.A. Gale, K.W. Church, and D. Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237.

M. Gerner, G. Nenadic, and C.M. Bergman. 2010a. An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In *BioNLP'10*, pages 72–80.

M. Gerner, G. Nenadic, and C.M. Bergman. 2010b. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85+.

M.A. Haendel, F. Neuhaus, D. Osumi-Sutherland, P.M. Mabee, J.L.V. Mejino, C.J. Mungall, and B. Smith. 2008. CARO–the common anatomy reference ontology. *Anatomy Ontologies for Bioinformatics*, pages 327–349.

M.A. Haendel, G.G. Gkoutos, S.E. Lewis, and C. Mungall. 2009. Uberon: towards a comprehensive multi-species anatomy ontology. *Nature precedings*.

J-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings JNLPBA'04*.

M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. 2008. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome biology*, 9(Suppl 2):S1.

A. Kumar, B. Smith, and D.D. Novotny. 2004. Biomedical informatics and granularity. *Comparative and functional genomics*, 5(6-7):501–508.

J.D. Lafferty, A. McCallum, and F.C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

N. Naderi, T. Kappler, C.J.O. Baker, and R. Witte. 2011. OrganismTagger: Detection, normalization, and grounding of organism entities in biomedical documents. *Bioinformatics*.

M. Neves, A. Damaschun, A. Kurtz, and U. Leser. 2012. Annotating and evaluating text for stem cell research. In *Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012)*. (to appear).

T Ohta, Y Tateisi, H Mima, and J Tsujii. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. *Proceedings of the Human Language Technology Conference (HLT 2002)*, pages 73–77.

N. Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

S. Pyysalo, T. Ohta, M. Miwa, H-C. Cho, J. Tsujii, and S. Ananiadou. 2012a. Event extraction across multiple levels of biological organization. (manuscript in review).

S. Pyysalo, T. Ohta, J. Tsujii, and S. Ananiadou. 2012b. Learning to classify anatomical entities using open biomedical ontologies. *Journal of Biomedical Semantics*. (to appear).

D. Rebholz-Schuhmann, A. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, K. Hornbostel, A. Kouznetsov, R. Witte, J. Laurila, C. Baker, C. Kuo, S. Clematide, F. Rinaldi, R. Farkas, G. Mora, K. Hara, L.I. Furlong, M. Rautschka, M. Neves, A. Pascual-Montano,

Q. Wei, N. Collier, M. Chowdhury, A. Lavelli, R. Berlanga, R. Morante, V. Van Asch, W. Daelemans, J. Marina, E. van Mulligen, J. Kors, and U. Hahn. 2011. Assessment of NER solutions against the first and second calbc silver standard corpus. *Journal of Biomedical Semantics*, 2(Suppl 5):S11.

C. Rosse and J.L.V. Mejino. 2003. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478–500.

C. Rosse and J.L.V. Mejino. 2008. The foundational model of anatomy ontology. *Anatomy Ontologies for Bioinformatics*, pages 59–117.

B. Smith, A. Kumar, W. Ceusters, and C. Rosse. 2005. On carcinomas and other pathological entities. *Comparative and functional genomics*, 6(7-8):379–387.

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S-A Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel, and S. Lewis. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the EACL 2012 Demonstrations*, pages 102–107.

L. Tanabe, N. Xie, L. Thom, W. Matten, and W.J. Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(Suppl 1):S3.

E.F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147.

R. Travillian, T. Adamusiak, T. Burdett, M. Gruenberger, J. Hancock, A-M. Mallon, J. Malone, P. Schofield, and H. Parkinson. 2011. Anatomy ontologies and potential users: bridging the gap. *Journal of Biomedical Semantics*, 2(Suppl 4):S3.

J. Wilbur, L. Smith, and L. Tanabe. 2007. BioCreative 2 Gene Mention Task. In *Proceedings of Second BioCreative Challenge Evaluation Workshop*, pages 7–16.