Anatomical Entity Recognition with Open Biomedical Ontologies

Sampo Pyysalo[*†] **Tomoko Ohta**[‡] **Jun'ichi Tsujii**[§] **Sophia Ananiadou**[*†]
[*]National Centre for Text Mining, Manchester, UK
[†]School of Computer Science, University of Manchester, Manchester, UK
[‡]Department of Computer Science, University of Tokyo, Tokyo, Japan
[§]Microsoft Research Asia, Beijing, China
sampo.pyysalo@gmail.com,okap@is.s.u-tokyo.ac.jp
jtsujii@microsoft.com,sophia.ananiadou@manchester.ac.uk

Abstract

Anatomical entities are central to much of biomedical discourse and must be considered in any attempt to fully analyse biomedical scientific text. However, while a wealth of tools and resources have been introduced in domain natural language processing efforts for the recognition of mentions of biomolecules and organisms in text, there has been little study of anatomical entities such as tissues and organs. In this paper, we consider the closely related tasks of distinguishing terms that refer to anatomical entities from ones that do not and assigning the former into upper-level semantic classes. We draw on the wealth of anatomical domain resources available in the Open Biomedical Ontologies repository, evaluating entity detection performance using the ontologies against a manually curated corpus of 5000 phrases occurring with high frequency in PubMed. We further analyse the upper-level structure of these ontologies to determine whether the resources offer a stable basis for species-independent classification and discuss remaining conflicts and challenges.

1 Introduction

The development of tools capable of automatically analysing natural language text to provide a structured representation of statements regarding the connections between molecular-level processes and organism-level effects is a long-standing goal of biomedical natural language processing (NLP) (Ananiadou et al., 2010). Mentions of anatomical entities such as *blood*, *epithelium* and *heart* are central to establishing such effects, yet their recognition in text has been largely neglected in recent biomedical NLP. A wealth of resources and systems have been introduced for the recognition of molecular entities such as genes/proteins (Kim et al., 2004; Settles, 2005; Hirschman et al., 2005; Krallinger et al., 2008; Leaman and Gonzalez, 2008) and chemicals (Corbett and Murray-Rust, 2006; Corbett et al., 2007; Nobata et al., 2010; Kolluru et al., 2011) and recent efforts have successfully addressed the recognition of organism mentions (Gerner et al., 2010a; Solt et al., 2010; Naderi et al., 2011). However, although the intermediate levels of biological organization are in scope of general-purpose term tagging tools (Aronson, 2001; Jonquet et al., 2009), anatomical entities have been specifically considered in few studies, with focus on a limited number of classes such as cells (Kim et al., 2004) or on mention detection without classification (Gerner et al., 2010b).

The definition of a task scope and semantic classes is a prerequisite for the systematic study of entity recognition. In this study, we take these first steps toward establishing a species-independent anatomical entity mention recognition task, aiming to maintain compatibility with existing ontologies and to complement established entity recognition tasks without overlap in scope. We curate a reference corpus of common anatomical entity mentions as they appear in text and study the use of ontologies for differentiating statements referring to anatomical entities from ones that do not (detection) and assigning a semantic class to each of the former (classification).

2 Resources

The resources presented in this section serve as the foundation for this work.

2.1 OBO anatomy resources

Following a preliminary review of available anatomical resources, we chose to base our study on the OBO (Open Biomedical Ontologies) resources (Smith et al., 2007). The OBO consortium seeks to develop orthogonal and mutually interoperative biomedical domain ontologies, and the OBO foundry website[1] currently lists 40 ontologies involving the domain "anatomy". These vary in scope from single species to classes of organisms or even a whole domain of life (Plant Ontology) and range in size from fewer than 100 to over 50,000 terms.

For this study, we initially analysed all anatomy ontologies in the OBO foundry to identify a set of resources that are non-redundant (including e.g. EHDAA2 but not its variants EHDA and EHDAA), define at least some physical anatomical entities (excluding e.g. FBdv, WBls, BSPO and ATO) of which a non-trivial part (over 5%) are organized in an **IS-A** hierarchy (excluding EMAP, MAT, and MFO) in order to allow the identification of the upper-level classes that specific anatomical entities belong to.[2] Through this procedure, we selected 26 anatomy ontologies (listed in Table 4), analysed and discussed further in the following.

The proliferation of species-specific ontologies is in part due to the lack of a species-independent ontology of anatomy comparable to the widely accepted OBO foundry ontologies for biological processes (GO-BP), cellular components (GO-CC) and cell types (CL). For broad coverage of anatomical terms, it is thus necessary to make combined use of the species-specific resources. However, combining a large number of ontologies with varying degrees of formality – many originally developed without reference to a commonly-accepted theory of anatomy or top-level ontology – holds substantial challenges. Two of the OBO resources, CARO and UBERON, seek to address these challenges.

[1] `http://www.obofoundry.org/`

[2] Details of this selection are available from the project page at `http://nactem.ac.uk`.

2.2 CARO

The Common Anatomy Reference Ontology (CARO) (Haendel et al., 2008) seeks to define a common basis for OBO anatomy resources. CARO defines a small upper-level ontology, less than 50 terms, based on the high-level structure of the extensive Foundational Model of Anatomy (FMA) ontology of human anatomy (Rosse and Mejino, 2003; Rosse and Mejino, 2008). CARO seeks to be applicable to all organisms and to capture the consensus of a broad group of investigators representing species-specific resources. There are ongoing efforts to standardize OBO anatomy ontologies on CARO through consolidation of upper-level ontology structures and the definition of explicit cross-references identifying identical terms.

CARO follows FMA in adhering to single inheritance and disjoint division of types. Thus, if each term in every OBO anatomy ontology were associated with exactly one CARO term, the upper-level structure of CARO could provide a unique, species-independent classification of any anatomical entity defined in the species-specific resources.

2.3 Uberon

The Uberon ontology (Haendel et al., 2009) aims to unify species-specific resources by combining them into a single multi-species ontology with explicit links to the various species-specific ontologies. Created through initial automatic alignment of anatomy resources and subsequent manual curation, Uberon currently defines 6,208 terms and includes 24,920 cross-reference links to other resources. The Uberon resources also include a "bridge" mapping that defines cross-references from other anatomy ontologies to Uberon.

Were Uberon to provide sufficient coverage of anatomical entity mentions, it could potentially eliminate the need to consider a large number of disparate resources to address the recognition task. However, Uberon defines its scope as metazoans – excluding e.g. bacteria and protozoa – and involves frequent multiple inheritance and non-disjoint types, factors that limit its coverage and complicate its use for assigning entities into non-overlapping classes.

Term	#
MATERIAL ANATOMICAL ENTITY	12
ANATOMICAL STRUCTURE	16
PORTION OF ORGANISM SUBSTANCE	14
IMMATERIAL ANATOMICAL ENTITY	12

Table 1: CARO top-level structure and number of other OBO anatomy ontologies defining each term. Indentation corresponds to **IS-A** structure.

3 Task definition

To define the task, we must first define *anatomical entity* and related key concepts. We propose to follow the definition of the candidate standard, CARO:

> ANATOMICAL ENTITY$_{\text{CARO}}$
> Biological entity that is either an individual member of a biological species or constitutes the structural organization of an individual member of a biological species.

CARO features a top-level division between material and immaterial anatomical entities, further dividing the former into structures and substances. Our analysis indicates that this division is adopted by approximately half of the 26 considered ontologies (Table 1); most of the others lack an explicit top-level structure.

CARO ANATOMICAL STRUCTURE subsumes most commonly recognized anatomical entities:

> ANATOMICAL STRUCTURE$_{\text{CARO}}$
> Material anatomical entity that has inherent 3D shape and is generated by coordinated expression of the organism's own genome.

This definition excludes e.g. pathological formations such as tumors (Smith et al., 2005) and simple chemicals such as carbon dioxide molecules (Rosse and Mejino, 2008).

CARO follows FMA in subdividing ANATOMICAL STRUCTURE primarily by granularity, along lines broadly corresponding to commonly recognized levels of biological organization (Kumar et al., 2004). This subdivision and the number of considered ontologies defining each term are shown in Table 2. We briefly note that while most of the terms are defined in roughly the same number of ontologies defining ANATOMICAL STRUCTURE, a few, in particular MULTI-CELL-COMPONENT-STRUCTURE, have very limited adoption in OBO resources.

Term	#
CELL COMPONENT	9
MULTI-CELL-COMPONENT STRUCTURE	4
ACELLULAR ANATOMICAL STRUCTURE	16
CELL	20
PORTION OF TISSUE	14
EXTRAEMBRYONIC STRUCTURE	13
MULTI-TISSUE STRUCTURE	13
COMPOUND ORGAN	12
ANATOMICAL GROUP	13
ORGANISM SUBDIVISION	15
MULTI-CELLULAR ORGANISM	14

Table 2: CARO ANATOMICAL STRUCTURE subdivision and number of other OBO anatomy ontologies defining each term.

3.1 Scope

Based on the preceding definitions, we propose the following scope for anatomical entity recognition:

> Mentions of anatomical structures, organism substances, and immaterial anatomical entities, excluding biological macromolecules and whole organisms.

Where biological macromolecules[3] and organisms[4] are excluded to avoid overlap with the established tasks of gene and gene product mention and organism mention recognition.

3.2 Classification

The many OBO anatomy ontologies are still far from full agreement on the upper-level ontological structure (Tables 1 and 2). Nevertheless, in providing a comprehensive, disjoint and species-independent upper-level structure with rough consensus support from domain resources, the CARO division is a strong candidate for a detailed, stable classification of anatomical entity mentions. We thus tentatively propose to adopt a 13-class classification using the 11 CARO subtypes of ANATOMICAL STRUCTURE and the disjoint classes PORTION OF ORGANISM SUBSTANCE and IMMATERIAL ANATOMICAL ENTITY.

[3] BIOLOGICAL MACROMOLECULE$_{\text{FMA}}$: Anatomical structure which has as its parts one or more ordered aggregates of nucleotide, amino acid fatty acid or sugar molecules bonded to one another.

[4] While we accept the view that unicellular organisms are indistinguishable from their cells, we *exclude* mentions of unicellular organism names as they are in scope of organism mention recognition.

4 Methods

We evaluate the feasibility of the detection task, study the coverage of the OBO anatomy ontologies, and evaluate the stability of the proposed classification using the methods presented in the following. As our goal is to gain insight into the task and resources, we avoid "black-box" machine learning methods and instead apply a simple approach that permits straightforward analysis.

4.1 Term matching

For determining whether a given candidate string refers to an anatomical entity or not, we attempt to match it against ontology terms. We assume a basic strategy where strings match only terms with a name or EXACT type synonym that is identical, also in case.[5] We further consider the following variants of this strategy:

Case-insensitive: match also if identical to a term name or synonym ignoring case. For example, "cns" matches "CNS".

Variants: match also if identical to a lexical variant of a term name or synonym as generated by the NLM Lexical Variant Generator.[6]

All synonyms: match also if identical to a synonym of a type other than EXACT.

Multiple terms can be retrieved, if, for example, a string matches the name of one term and a synonym of another.

4.2 Anatomical entity classification

The clear majority of the OBO anatomy ontologies adopt an **IS-A** hierarchy as their primary organization. This permits a simple approach to classification: given a term matching an input string, trace **IS-A** links until a term of the desired level of generality is found, and return the name of that term. The assigned classes are thus simply the names of upper-level terms. This approach is illustrated in Figure 1.

In practice, this method often fails to resolve to exactly one class. Not all of the resources define **IS-A** links for all their terms, and many can define more than one per term (multiple inheritance), leading to several candidate classes. This issue can arise also from multiple matches for an input string. In all cases, we retrieve the full set of unique relevant upper-level terms[7] for each ontology in which string matches were found.

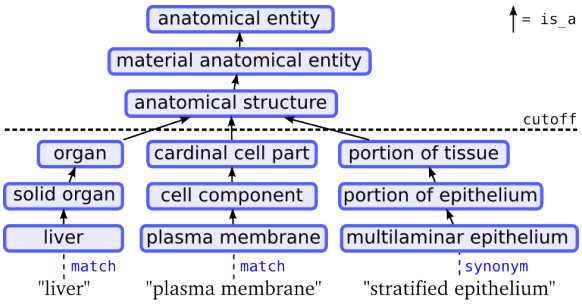

Figure 1: Idealized illustration of classification approach. Dotted line marks desired level of generality. Example simplified from FMA.

4.3 Ontology mapping

For deciding whether candidate classes assigned on the basis of different ontologies agree, we determine whether their corresponding ontology terms are equal. We consider the following approaches:

Name match: two terms are equal if their names match under case-insensitive string matching.

CARO mapping: two terms are equal if they map to the same CARO term, either through explicit cross-reference or through matching either in name or synonym.

Uberon mapping: two terms are equal if they map to the same Uberon term, either through cross-reference or through links in Uberon bridge.

These approaches are compared through their effect on the consistency and ambiguity of the resulting top-level classification.

5 Data and annotation

As there is, to the best of our knowledge, no established species-independent corpus annotated for anatomical entity mentions, we created a new reference corpus for evaluation. Candidate anatomical entity mentions were selected on the basis of simple overall mention frequency in PubMed

[5] We normalize case in ontologies throughout, converting e.g. FMA "Plasma membrane" and SAO "Plasma Membrane" into "plasma membrane".

[6] http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/

[7] Relevance cutoff levels were determined individually for each ontology below terms corresponding to ANATOMICAL ENTITY$_{CARO}$ and ANATOMICAL STRUCTURE$_{CARO}$.

Class	#	
Organism	334	6.7%
Anatomical entity	409	8.2%
Biological macromolecule	543	10.9%
Total	5000	100%

Table 3: Annotation statistics

to avoid biasing the sample toward e.g. species, scale/granularity, or subdomains of biomedicine.

We first selected a random sample of 200,000 PubMed citations from the PubMed 2011 distribution, corresponding to approximately 1% of all citations. The titles and abstracts of these documents were then extracted and split into sentences using the GENIA sentence splitter[8] and the syntactic structure of the sentences analysed with the Charniak and Johnson (2005) parser[9] using the self-trained biomedical model of McClosky (2009). This processing resulted in a corpus of 924,036 sentences containing a total of 8,188,974 noun phrases (NPs).

We then normalized the NPs by removing determiners and lowercasing for sentence-initial capitalization, removed NPs of fewer than three letters (as excessively ambiguous) and determined the number of documents in the sample in which each normalized phrase appears. The NPs were then ordered by this frequency of appearance, and the highest-ranking 5000 were manually examined by a PhD biologist to identify phrases which in at least one of their frequent senses refer to an anatomical entity, also separately marking references to biological macromolecules (under the FMA definition) and whole organisms. As there are no established criteria for assigning anatomical entities to specific classes – indeed, as this study aims to determine whether some stable assignment can be performed using OBO resources – no detailed classification was performed in the annotation.

Table 3 shows the statistics of the resulting corpus. Anatomical entities are referenced by these common phrases with roughly comparable frequency to biological macromolecules and organisms, providing a quantitative confirmation of their central role in biomedical scientific text.

[8] http://www-tsujii.is.s.u-tokyo.ac.jp/~y-matsu/geniass/

[9] https://github.com/dmcc/bllip-parser

6 Experiments

6.1 Experimental setup

We report the results of anatomical entity mention detection using the standard precision and recall metrics, summarizing results for the two as F-score.[10] For evaluating the consistency of the classification provided by the ontologies, we report the average number of ontologies using which a class could be assigned ("matches") and the average number of distinct classes ("classes") assigned for each phrase. To summarize these results, we calculate the match/class ratio. Ideally, only a single class would be assigned to each phrase and this assignment would be supported by multiple ontologies, giving a high match/class ratio. Conversely, if no two resources ever agree on class assignment, the ratio will be 1 (assuming single inheritance; under multiple inheritance values below 1 are possible).

Throughout the experiments we applied filtering to ignore matches in branches of ontologies not relating to anatomy, for example ignoring all terms in the GO BIOLOGICAL PROCESS subontology. We similarly excluded as out of scope for the targeted definition all mentions of biological macromolecules and whole organisms, excluding for example the GO PROTEIN COMPLEX and the FMA BIOLOGICAL MACROMOLECULE branch.

The manually annotated corpus was only used in the final experiments, with all development performed with reference to a small separate dataset.

7 Results

7.1 Mention detection

Table 4 summarizes the results for anatomical entity mention detection. While precision is high in almost all cases, there are enormous differences in recall between the ontologies, with the majority of the species-specific resources providing less than 10% recall of commonly appearing anatomical terms, while a small number achieve over 50% recall standalone. The overall trend agrees with expected species biases inherent to PubMed, with e.g. human and mouse resources (FMA, MA) ranking high and fungal and slime mold ontolo-

[10] $F_1 = \frac{2pr}{p+r}$, where p is precision and r recall. We use "F-score" for F_1 score throughout.

Ontology	Basic	Case-ins.	Variants	All-synon.
FMA	90/46/60.8	90/46/60.8	91/67/**76.8**	90/46/60.8
BTO	94/35/51.3	94/35/51.3	95/53/**68.2**	92/42/57.5
UBERON	95/43/59.4	95/44/59.7	94/51/**66.3**	92/46/61.5
MA	99/31/47.7	99/31/47.7	99/38/**55.3**	99/34/50.9
ZFA	88/21/33.5	82/21/33.5	90/28/**42.6**	83/21/33.5
TAO	76/18/28.5	74/18/28.7	82/29/**42.4**	76/18/28.5
XAO	99/17/29.1	99/17/29.1	99/24/**38.9**	99/19/32.2
EHDAA2	100/15/26.2	100/15/26.2	100/19/**32.3**	100/15/26.2
FBbt	89/8/14.1	86/8/14.1	85/13/**22.3**	86/9/16.9
CL	75/0.7/1.5	75/0.7/1.5	98/11/**20.0**	42/0.7/1.5
AAO	90/6/12.0	90/6/12.0	92/11/**19.4**	90/7/12.4
GO	75/4/8.4	76/5/8.8	81/9/**15.6**	58/4/8.3
HAO	73/5/10.1	73/5/10.1	65/8/**14.1**	74/8/13.9
SAO	67/2/4.8	67/2/4.8	83/7/**13.6**	67/2/4.8
TGMA	86/3/5.7	86/3/5.7	83/5/8.9	60/8/**13.6**
WBbt	75/4/7.1	71/4/7.0	83/6/**11.5**	77/6/10.6
BILA	94/4/7.6	94/4/7.6	92/6/**10.7**	95/4/8.5
AEO	100/2/4.8	100/2/4.8	100/4/**7.6**	100/3/5.3
PO	100/2/3.4	100/2/3.4	100/3/**6.2**	90/2/4.3
DC_CL	100/2/3.9	100/2/3.9	100/3/**5.8**	100/2/3.9
VAO	88/2/3.4	88/2/3.4	86/3/**5.7**	88/2/3.4
SPD	90/2/4.3	90/2/4.3	86/3/**5.7**	90/2/4.3
FAO	100/0.5/1.0	100/0.5/1.0	100/1/**2.0**	100/0.5/1.0
CARO	100/0.5/1.0	100/0.5/1.0	100/1/**2.0**	100/0.5/1.0
TADS	100/0.5/1.0	100/0.5/1.0	100/1/**2.0**	100/0.5/1.0
DDANAT	50/0.2/0.5	50/0.2/0.5	60/0.7/**1.5**	33/0.2/0.5
ALL	78/58/66.9	76/58/66.2	78/80/**79.2**	72/60/65.4

Table 4: Anatomical entity mention detection results (precision/recall/F-score)

gies (FAO, DDANAT) having low coverage. Interestingly, the highest standalone performance is achieved through the use of the human-specific FMA, not the multi-species Uberon. Nevertheless, overall highest F-score results are achieved through combined use of all the resources (ALL).

The case-insensitive and all-synonyms matching strategies show mixed results, including a negative overall effect on the combination using all of the ontologies. By contrast, the variant matching strategy shows a consistent positive effect, including an over 10% point F-score improvement for the combination. Further combinations of the matching strategies did not improve on this result (data not shown).

As the best overall result, we find that anatomical entity mentions can be distinguished from other common phrases with nearly 80% F-score and balanced precision and recall using an approach relying only on string matching against OBO anatomy term names and synonyms and their lexical variants. This indicates that the detection task is well-defined and feasible, and suggests that a high level of detection reliability might be achievable using more sophisticated methods.

Mapping	Matches	Classes	Ratio
Name match	3.54	3.15	1.12
CARO mapping	3.56	2.96	1.20
Uberon mapping	4.71	2.39	1.97

Table 5: Anatomical entity classification results: average number of ontologies through which a class could be assigned, number of unique classes, and their ratio.

7.2 Classification

For class comparison, we classified all of the gold anatomy phrases using the best settings from mention detection experiments (ALL+Variants). Table 5 shows results for the experiments evaluating the consistency of the classification provided by the ontologies, using the approaches described in Section 4.3 for determining whether terms in different ontologies agree.

As expected, there is substantial overlap between the ontologies: on average, a class can be assigned to a phrase on the basis of more than three different ontologies. However, this overlap reveals a striking frequency of disparities in the upper-level classes entities are assigned to, with the average number of different classes nearly matching the number of ontologies. Mapping to CARO improves the agreement only slightly, while bridging to Uberon has a more substantial effect, bringing the average number of ontologies supporting the assignment of each candidate class close to two. However, even this effect is primarily due not to elimination of ambiguity, but rather to an increase in the number of ontologies through which a class can be assigned.[11]

These results suggest that despite unification efforts, the OBO anatomy resources disagree on the upper-level class of many of even the most frequently discussed anatomical entities, a finding that calls into question whether the resources can provide a stable basis for consistent organism-independent anatomical entity classification. In light of this result, we performed a more detailed analysis of the classification, described in the following section.

[11] This somewhat unintuitive effect of applying the Uberon bridge is explained by the presence of ontologies that have incomplete IS-A structure: without mapping to another ontology, no upper-level class can be identified for many strings matched in these ontologies.

Term	Classes
brain	CARDINAL ORGAN PART (FMA), MULTI-TISSUE STRUCTURE (FBbt), COMPOUND ORGAN (ZFA,TAO), HEAD ORGAN (MA), UNCLASSIFIED (AAO),
peripheral nervous system	SET OF ORGANS (FMA), ANATOMICAL SYSTEM (ZFA,TAO), ORGAN SYSTEM SUBDIVISION (FBbt)
nerve	PORTION OF TISSUE (AAO,XAO,EHDAA2,ZFA,TAO,AEO), MULTI-CELL-COMPONENT STRUCTURE (FBbt), CARDINAL ORGAN PART (FMA)
blood vessel	MULTI-TISSUE STRUCTURE (ZFA,TAO), PORTION OF TISSUE (EHDAA2,AEO)
sternum	MULTI-TISSUE STRUCTURE (EHDAA2), ORGAN SYSTEM SUBDIVISION (FBbt), COMPOUND ORGAN (FMA), AREA (HAO)
ganglion	PORTION OF TISSUE (AAO,XAO,EHDAA2,ZFA,TAO,AEO), MULTI-TISSUE STRUCTURE (FBbt,FMA)
mesoderm	EMBRYONIC STRUCTURE (BILA,AAO), PORTION OF TISSUE (ZFA,TAO), DEVELOPING ANATOMICAL STRUCTURE (EHDAA2), GESTATIONAL STRUCTURE (FMA)
blastomere	EMBRYONIC STRUCTURE (BILA,TAO), CELL (AAO,XAO,ZFA), GESTATIONAL STRUCTURE (FMA)

Table 6: Disagreements in upper-level classification for some common anatomical entities.

7.3 Analysis and discussion

To get a more detailed understanding of the classification disagreements, we calculated a confusion matrix by determining for each pair of classes the number of gold anatomical entities assigned both classes when performing classification using all the ontologies.

Analysis of the confusion matrix indicated that within CARO, CELL, PORTION OF TISSUE, ANATOMICAL GROUP and ORGANISM SUBDIVISION were assigned relatively frequently and consistently, each with over 100 entity mentions matching in multiple ontologies assigned the class, and 70% or more of such cases agreeing on the assignment. By contrast, there was frequent disagreement in the "middle granularity" range of MULTI-TISSUE STRUCTURE and COMPOUND ORGAN, where the high-coverage ontologies FMA, MA and UBERON define the additional non-CARO term ORGAN. The other CARO classes were assigned in only relatively few cases.

Table 6 illustrates a number of cases where the ontologies disagree on upper-level classification, showing also non-CARO terms. Anatomical structures relating to development (GESTATIONAL STRUCTURE / DEVELOPING STRUCTURE / EMBRYONIC STRUCTURE) were a particularly frequent source of disagreement involving non-CARO terms. Further, the disagreements are not, in general, limited to cases that could be argued to be actual species differences and in many cases involve also prominent OBO ontologies (e.g. disagreements between FMA and the "foundry" ontology ZFA), suggesting that many reflect genuine disagreement on foundational principles.

8 Conclusions

In this study, we have taken the first steps toward establishing an organism-independent anatomical entity recognition task, proposing a task scope and a tentative set of detailed entity classes based on the Common Anatomy Reference Ontology (CARO), introduced a corpus of 5000 common phrases from the biomedical scientific literature manually annotated for anatomical entity detection, and presented initial experiments evaluating the feasibility of the detection and classification tasks using 26 anatomy ontologies from the OBO (Open Biomedical Ontologies) foundry.

Experiments showed that anatomical entity mentions can be differentiated from other common phrases with nearly 80% precision and recall using a simple strategy based on ontology term matching, indicating that the entity detection task is both well-defined and feasible. However, the OBO resources were found to frequently disagree on the upper-level classification of anatomical entities, an issue that analysis indicated to be particularly common for multi-tissue structures, compound organs, and development-related structures. The resolution of the challenges in establishing a stable, detailed, organism-independent upper-level classification of anatomical entities remains future work.

All the resources and tools introduced in this study, including the manually annotated corpus, the detailed results of the analysis of classification disagreements, and the anatomical entity detection and classification system, are freely available from `http://nactem.ac.uk`.

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). The UK National Centre for Text Mining is funded by UK Joint Information Systems Committee (JISC).

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. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of AMIA*, pages 17–21.

E. Charniak and M. Johnson. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *ACL'05*, pages 173–180.

P. Corbett and P. Murray-Rust. 2006. High-throughput identification of chemistry in life science texts. *Computational Life Sciences II*, pages 107–118.

P. Corbett, C. Batchelor, and S. Teufel. 2007. Annotation of chemical named entities. In *BioNLP'07*, pages 57–64.

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

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

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*.

L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl.1)(suppl. 1):S1.

C. Jonquet, N.H. Shah, and M.A. Musen. 2009. The open biomedical annotator. *Summit on Translational Bioinformatics*, 2009:56.

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*.

B. Kolluru, L. Hawizy, P. Murray-Rust, J. Tsujii, and S. Ananiadou. 2011. Using workflows to explore and optimise named entity recognition for chemistry. *PloS one*, 6(5):e20181.

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.

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

R. Leaman and G. Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

D. McClosky. 2009. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Department of Computer Science, Brown University.

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

C. Nobata, P.D. Dobson, S.A. Iqbal, P. Mendes, J. Tsujii, D.B. Kell, and S. Ananiadou. 2010. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, pages 1–8.

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. Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

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.

I. Solt, D. Tikk, and U. Leser. 2010. Species identification for gene name normalization. *BMC Bioinformatics*, 11:1–2.