A New Corpus to Support Text Mining for the Curation of Metabolites in the ChEBI Database

Matthew Shardlow[1], **Nhung Nguyen**[1], **Gareth Owen**[2], **Steve Turner**[2], **Claire O'Donovan**[2], **Andrew Leach**[2], **John McNaught**[1], **Sophia Ananiadou**[1]
National Centre for Text Mining, University of Manchester[1], European Bioinformatics Institute[2]
{matthew.shardlow, nhung.nguyen, john.mcnaught, sophia.ananiadou}@manchester.ac.uk,
{gowen, turner, odonovan, arl}@ebi.ac.uk

Abstract

We present a new corpus of 200 abstracts and 100 full text papers which have been annotated with named entities and relations in the biomedical domain as part of the OpenMinTeD project. This corpus facilitates the goal in OpenMinTeD of making text and data mining accessible to the users who need it most. We describe the process we took to annotate the corpus with entities (*Metabolite*, *Chemical*, *Protein*, *Species*, *Biological Activity* and *Spectral Data*) and relations (*Isolated From*, *Associated With*, *Binds With* and *Metabolite Of*). We report inter-annotator agreement (using F-score) for entities of between 0.796 and 0.892 using a strict matching protocol and between 0.875 and 0.963 using a relaxed matching protocol. For relations we report inter annotator agreement of between 0.591 and 0.693 using a strict matching protocol and between 0.744 and 0.793 using a relaxed matching protocol. We describe how this corpus can be used within ChEBI to facilitate text and data mining and how the integration of this work with the OpenMinTeD text and data mining platform will aid curation of ChEBI and other biomedical databases.

Keywords: Text Mining, Corpus, Bioinformatics

1. Introduction and Background

The ChEBI (Chemical Entities of Biological Interest) database (Degtyarenko et al., 2008; Hastings et al., 2016) is a freely available, electronic dictionary and ontology of small molecules. ChEBI was created to help researchers in the field of molecular biology who need to know the structure, names, and properties of the small molecules that they encounter in their research. There are a number of freely-available chemical databases. Most of them are created by an automatic 'pipeline' process and contain information on polymers, industrial chemicals, synthetic intermediates, etc. Their sheer size creates problems for users, as any search may result in hundreds or even thousands of answers. For non-expert users it is very difficult to determine which, if any, is the compound they are really looking for. By contrast, the focus of the ChEBI database is on high quality rather than quantity.

ChEBI is manually curated and focuses on the requirements of the molecular biology community. Manual curation assures the high quality of the database, but it also makes it an expensive database to produce, particularly so for our focus of metabolites. Users who would like to add a new metabolite, may only know a research code or a trivial name from which it is not possible to deduce a structure. The curator then has to search through scientific literature to find as much information as possible about the new metabolite. The information is likely to include:

- In which species is it present?
- Does it have any interesting biological properties, applications, etc? (E.g., Biological activity)
- Is there any spectral data available that indicates the structure?
- From which chemical's metabolism does it derive?

All of this information ideally needs to be supported by appropriate citations to publications in the scientific literature. Whilst we cannot automate every step in the curation process, we can use text mining to facilitate the curator throughout the process.

This work is part of the OpenMinTeD project[1] (Oudenhoven and Pontika, 2017), which is developing a text and data mining framework consisting of an integrated registry, metadata schema, text mining workflow service and annotation viewer. The OpenMinTeD project works with data providers to give users access to large databases of open access publications (Knoth and Pontika, 2016). The OpenMinTeD project is also promoting a number of community standards for text and data mining (Przybyła et al., 2016; Ba and Bossy, 2016; Peters, 2016), such as UIMA (Ferrucci and Lally, 2004), maven[2] and Docker (Merkel, 2014), which we have adopted in our work. The OpenMinTeD project has also made extensive efforts to facilitate the legal interoperability of text mining software and content. From which, we have adopted lessons and advice for our work (Margoni and Dore, 2016; Labropoulou et al., 2016)

To facilitate curation in ChEBI through text mining, we will deploy this corpus and any tools derived from it via the OpenMinTeD platform. The platform provides many advantages to both the data provider and consumer, including persistent availability, support and training, sustainability of the platform, reusability of tools and redeployability of tools to new scenarios. In addition, the tools we make available will be usable by audiences other than our intended target (curators of ChEBI), fulfilling the project's goal of promoting open science throughout communities which stand to benefit from text and data mining.

The intended purpose of this corpus is to provide training

[1] `https://openminted.eu`

[2] `https://maven.apache.org/`

data for both named entity recognition and relation extraction tools. We briefly describe approaches to each of these areas below.

Named entity recognition (Nadeau and Sekine, 2007) is a common text mining task in which an algorithm is used to identify parts of an unstructured text that can be categorised according to a given schema. Named entity recognition has been applied across diverse corpora including but not limited to Twitter (Derczynski et al., 2015; Baldwin et al., 2015), biomedical papers (Munkhdalai et al., 2015; Wang et al., 2015) and other areas. Approaches for named entity recognition range from using dictionaries (Cohen and Sarawagi, 2004) to using regular expressions and other rules (Kluegl et al., 2016; Chiticariu et al., 2010), using machine learning approaches such as the conditional random field (Lu et al., 2015), or more recently leveraging advances in the field of deep learning (Chiu and Nichols, 2015; Lample et al., 2016).

Relation extraction (Aggarwal and Zhai, 2012) is a task in information extraction that requires an algorithm to link together two named entities according to a given schema which defines the meaning of the link. Relation extraction requires named entity recognition to be performed as a prerequisite. Relation extraction may be performed using supervised (Nguyen and Grishman, 2015; Pons et al., 2015) or unsupervised methods (Quan et al., 2014). Simple heuristics such as two entities in close proximity may serve as a baseline, but are less powerful than measures which take structural and context features of the named entities into account.

2. Corpus Construction

We annotated a set of 200 abstracts and 100 full papers with the entities that were of interest to us, as well as relations between these entities. Although we could have leveraged existing corpora for some of the entity types that we were interested in, we found that there were very few corpora dealing with metabolites in the fine-grained annotations that we were interested in. In addition, whereas corpora may have contained the entity types of interest to us, no suitable corpora existed containing the relation types we wished to annotate.

We first defined the entity types that were of interest to us. These were determined as follows. The definitions of these entities were driven by the curators of ChEBI, who also participated in the annotation process.

Metabolite: A chemical which has been produced by, detected in, or isolated from a living organism, where this is clear from the context of the paper (e.g. Nitrosobenzene, 11-Deoxycorticosterone, sclerotiorin).

Chemical: Any name that is used to define 'small' chemicals (those that are not proteins, nucleic acids, etc.). Includes molecules, salts, class names (e.g. benzoate esters; indole alkaloids; etc.) and groups (parts of molecules) - e.g. methyl group; benzyl substituent; alanine residue. . .)

Protein: Any protein or large polypeptide (usually one that is too big to be drawn by normal chemical drawing software). All enzymes and receptors are considered to be proteins in our scheme (e.g. 4-Dihydroxyphenylalanine decarboxylase, Dopa decarboxylase).

Species: Any entity referring to a formal name for a living organism or from which the name can be inferred (e.g. 'volunteer', 'patient' implies 'human').

Biological Activity: An effect/consequence that a chemical entity has on a biological system. Examples may include affecting the activity (e.g. by inhibition or activation) of a particular enzyme; growth regulator; antimicrobial agent; apoptosis inducer; anti-inflammatory; flavour enhancer; etc.

Spectral Data: data arising from spectrometry. e.g., 1H-NMR, 13C-NMR, MS, X-ray, IR, etc. Where two or more spectroscopic techniques are present, each should be tagged separately.

We also defined the following types of relation between the entities. The relations help us to answer the questions about a new metabolite that a curator may have, as outlined in Section 1..

Isolated_from(Metabolite, Species): A metabolite was isolated from or detected in a specific species.

Associated_with(Chemical / Metabolite, Biological Activity / Spectral Data): A chemical or metabolite is linked to a particular biological activity or spectral data.

Binds_With(Chemical / Metabolite, Protein): A chemical or metabolite interacts with (e.g.binds) and affects the behaviour of a biological target.

Metabolite of(Metabolite, Chemical): A metabolite is derived from the metabolism of a related compound.

We developed guidelines containing these definitions. The guidelines provided examples of the annotations, as well as specific information for each category that helped the annotators to agree in ambiguous cases. The guidelines were updated after each round of annotation in accordance with feedback from the annotators.

We selected 200 abstracts for annotation from PubMed, according to the criteria that each abstract should contain at least one of the relation types that we were looking for. We performed double annotation for all 200 abstracts to ensure the consistency and validity of our annotations. We engaged two annotators, both of whom were actively involved in the curation of the ChEBI database. We initially annotated 20 abstracts and evaluated inter-annotator agreement. We found some discrepancies between the annotators' interpretations of the categories and so we discussed these with the annotators and updated our definitions and the guidelines accordingly. We proceeded to annotate a further 20 abstracts, after which we obtained a higher agreement on all categories. We performed further resolution between the annotators and updated the guidelines, with further rounds of 60 and then 100 abstracts to bring the total number of doubly annotated abstracts up to 200.

Following on from this large round of double annotation, we proceeded to singly annotate the full text of 100 papers. It has been shown elsewhere in the literature that the annotation of full texts is preferential over the annotation of abstracts (Westergaard et al., 2017). We extracted full texts using the API provided by the Elsevier developer portal.[3] We chose 100 papers for annotation, again according to the criteria that they should contain at least one relation type of interest to us. Both annotators contributed to these annotations, although each full paper was only annotated by one annotator.

Both the corpus and guidelines associated with this paper will be made available via the OpenMinTeD platform, as well as in the supplementary material to this work, upon final submission.

3. Corpus Statistics

We calculated inter annotator agreement using the F-score statistic, as is the norm for text mining tasks. We could not use the more common Kappa statistic, as this takes the true negative rate into account, which is not appropriate for named entity recognition where all tokens or spans that have not been annotated will be considered true negatives. We calculate the F-score using both a strict and relaxed matching protocol as described below. The matching protocols are further explained in Table 1.

In the strict matching protocol, the annotators are considered to agree on a named entity only if they have annotated the exact same span and assigned the same entity type. In this case, the annotation will be considered a true positive. To obtain the false positive and false negative rate, we consider one annotator to be the 'gold standard'. A false negative is assigned if the gold annotator has created an annotation that is not present in the other annotations. A false positive is assigned if the gold annotator has not created an annotation that is present in the other annotations. We follow the same protocol for relations, where a true positive is assigned if both terms match exactly, as well as the type of the relation.

In the relaxed matching protocol, the annotators are considered to agree on a named entity only if the spans of two entities overlap by at least one character and the category is the same. False positives and false negatives are assigned as above when a matching term cannot be found for either the gold standard or the other annotator. Relations are assigned as a true positive if the terms match using the relaxed matching criteria and the category is the same.

The strict matching protocol may be overly punitive in cases where the annotators clearly agree on an annotation, but disagree about exactly which part of a term should be covered by the annotation. For example, consider the annotation *'poly aromatic hydrocarbons'*. One annotator may annotate the whole span as a chemical, whereas the other annotates the span *'aromatic hydrocarbons'*. In this case, the annotators both agree that there is a chemical at this point, the disagreement is around whether to include the term 'poly' in the annotation or not. According to the strict matching protocol this would be regarded as a false negative for the first annotator and a false positive for the second. According to the relaxed matching criteria, this would be considered a true positive. The relaxed matching criteria may be overly lenient, as even one character in common could signify a true positive. However, we hope that by providing both measures the reader will have the best tools to interpret our results. We have provided our results for inter annotator agreement in Table 2.

We can see from the results in Table 2 that we were able to attain a high level of agreement for entities between the annotators on the set of 200 abstracts in our corpus. It is clear to see that using the relaxed matching protocol yields an increase in agreement over the strict matching protocol. For the entities, the largest increase is for *Spectral Data*, where using relaxed matching gives an increase of 0.122, indicating that the annotators often disagree about the exact boundaries of *Spectral Data*. The *Species* category was the named entity with the smallest increase when using relaxed matching of only 0.05, indicating that the annotators rarely disagreed on the exact boundaries of each *Species* annotation. Overall, the highest agreement was attained for the *Species* annotation when using the strict matching protocol and for *Spectral Data* when using the relaxed matching protocol.

The agreement for relations is lower than for entities. Each relation covers two entities, so there is greater scope for the annotators to disagree on how these entities should be related. Furthermore, if the annotators have not agreed on the scope of one or both entities in the relation then it will not be considered a match according to the strict protocol. Accordingly, we can see that using the relaxed protocol for relations generally gives a larger performance boost than for entities alone. The largest increase is for the *Associated With* category, where relaxed matching yields an increase of 0.197. The *Associated With* category covers the *Spectral Data* entity which had the largest increase of all entities when using relaxed matching vs. strict, which may explain the increase for *Associated With*. The smallest increase within the relations is for *Binds With* where an increase of only 0.051 is recorded. This covers the *Metabolite*, *Chemical* and *Protein* entities, all of which had similarly small increases. Overall, the agreement for relations under relaxed matching gives scores between 0.744 for *Binds With* and 0.793 for *Associated With*. These figures indicate that agreement is high, if not perfect, and that annotators may not always agree on the exact boundaries of each annotation.

The final corpus contains 200 doubly annotated abstracts and 100 singly annotated full papers. The full papers portion of the corpus is much longer than that of the abstracts portion as each full paper is much longer than a single abstract. We present statistics on the full corpus in Table 3.

We can see from Table 3 that the full papers are indeed much richer than the abstracts. For entities, the average full paper contain between 12.08 (*Species*) and 28.57 (*Spectral Data*) times more entities than the average abstract. The increase in data is lower for relations, where the average full paper contains between 2.32 (*Metabolite Of*) and 8.09 (*Associated With*) times more relations than the average ab-

[3] https://dev.elsevier.com/

	$\underbrace{\overbrace{\text{benzoate esters}}^{\text{annotator 1}}}_{\text{annotator 2}}$ are	$\underbrace{\overbrace{\text{benzoate ester}}^{\text{annotator 1}}\text{s}}_{\text{annotator 2}}$ are	$\overbrace{\text{benzoate esters}}^{\text{annotator 1}}$ are
Strict	Match	No Match	No Match
Relaxed	Match	Match	No Match

Table 1: Three possible annotation scenarios and the results of the strict and relaxed matching protocol. The first column shows an ‘exact match’ where both annotators have highlighted the same term. This is considered a match by both protocols. The second column shows a partial match, where the annotators have agreed on the annotation, but disagreed on the scope. The strict matching protocol does not consider this a match, whereas the relaxed matching protocol does consider this a match. The final column shows that annotator 1 made an annotation, where annotator 2 did not. Neither protocol would consider this case to be a match.

Category	Strict	Relaxed
Metabolite	0.821	0.875
Chemical	0.865	0.950
Protein	0.866	0.944
Species	0.892	0.942
Biological Activity	0.796	0.904
Spectral Data	0.841	0.963
Isolated From	0.591	0.766
Associated with	0.596	0.793
Binds With	0.693	0.744
Metabolite Of	0.623	0.789

Table 2: The agreement between annotators on entities (top) and relations (bottom) using strict and relaxed matching protocols. All reported values are F-score.

Category	Abstracts	Full Papers	Scale
Documents	200	100	—
Words	160.34	3722.80	23.22
Sentences	10.37	112.15	10.81
Metabolite	3.30	48.45	14.68
Chemical	14.49	213.29	14.72
Protein	4.14	63.01	15.22
Species	2.72	32.87	12.08
Biological Activity	2.98	47.51	15.94
Spectral Data	0.28	8.00	28.57
Isolated From	1.29	8.00	6.20
Associated With	4.45	36.01	8.09
Binds With	1.32	6.18	4.68
Metabolite Of	0.75	1.74	2.32

Table 3: The averaged statistics per document for both abstracts and full papers in our corpus. The first line of data shows the total number of each document type. It is clear that the full papers contain much more data than the abstracts alone. We present general stats (top), entities (middle) and relations (bottom). The third column shows the magnitude of the increase in number of available entities when using full texts as opposed to abstracts.

stract. The lower increase in relations, may be because they are more difficult for the annotator to spot in full papers, where a relation may span several paragraphs.

It is interesting to note that whilst the number of words in a full paper is on average 23.22 times greater than in abstracts, most of the increases for entities and relations are below this number (all except for *Spectral Data*). This demonstrates that whilst full papers are richer in availability of entities and relations, they are less densely packed with entities and relations than abstracts. This implies that a larger volume of entities and relations could be found by processing the same number of words from abstracts than from full papers. However, full papers (where available) provide much more information about the entities they contain and so are important for information extraction tasks where important information may not be reported in the abstract.

The *Chemical* entity is the most frequently reported entity in both the abstracts and full papers in our corpus, whereas *Spectral Data* is the the least frequent. The documents we have chosen are from biomedical journals and so it is unsurprising that they mention chemicals with such high frequency. *Spectral Data* is surprisingly low in abstracts, indicating that authors do not commonly report this entity in the abstract, but instead report it in the full text of the paper.

The *Associated With* relation is the most frequent relation in both the abstracts and full papers, whereas the *Metabolite Of* relation is the least frequent. *Associated With* is a broad category that covers several entity types, which may explain why it is more frequent than the other more narrowly scoped relations.

4. Conclusion

We have described our new corpus containing 200 abstracts and 100 full papers annotated with entities and relations that are useful for automating the curation process of the ChEBI database. This corpus will be made available as part of the OpenMinTeD project for use in text mining applications. We will use the corpus to train text mining tools capable of detecting the entities and relations contained within the corpus. These text mining tools will be made available via the OpenMinTeD platform for use in the curation of ChEBI, as well as for use by other teams of curators who face similar problems in automating their curation workflows.

5. Bibliographical References

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data.* Springer Science & Business Media. Section 2.3.

Ba, M. and Bossy, R. (2016). Interoperability of corpus processing work-flow engines: the case of alvisnlp/ml in openminted. In Richard Eckart de Castilho, et al., editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 15–18, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Baldwin, T., de Marneffe, M.-C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., and Vaithyanathan, S. (2010). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1002–1012, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chiu, J. P. C. and Nichols, E. (2015). Named entity recognition with bidirectional LSTM-CNNs. *CoRR*, abs/1511.08308.

Cohen, W. W. and Sarawagi, S. (2004). Exploiting dictionaries in named entity extraction: Combining semi-markov extraction processes and data integration methods. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 89–98, New York, NY, USA. ACM.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., AlcÃ¡ntara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl1):D344–D350.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Ferrucci, D. and Lally, A. (2004). Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214–D1219.

Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., and Puppe, F. (2016). Uima Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40.

Knoth, P. and Pontika, N. (2016). Aggregating research papers from publishers' systems to support text and data mining: Deliberate lack of interoperability or not? In Richard Eckart de Castilho, et al., editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 1–4, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Labropoulou, P., Piperidis, S., and Margoni, T. (2016). Legal interoperability is-sues in the framework of the openminted project: a methodological overview. In Richard Eckart de Castilho, et al., editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 60–63, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Lu, Y., Ji, D., Yao, X., Wei, X., and Liang, X. (2015). CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *Journal of Cheminformatics*, 7(1):S4, Jan.

Margoni, T. and Dore, G. (2016). Why we need a text and data mining exception (but it is not enough). In Richard Eckart de Castilho, et al., editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 57–59, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239), March.

Munkhdalai, T., Li, M., Batsuren, K., Park, H. A., Choi, N. H., and Ryu, K. H. (2015). Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *Journal of Cheminformatics*, 7(1):S9.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nguyen, T. H. and Grishman, R. (2015). Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado, June. Association for Computational Linguistics.

Oudenhoven, M. and Pontika, N. (2017). Learning about text and data mining: The future of open science. In *Proccedings of the Open Science Conference*, Berlin, Germany.

Peters, W. (2016). Tackling resource interoperability: Principles, strategies and models. In Richard Eckart de Castilho, et al., editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 34–37, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Pons, E., Becker, B., Akhondi, S., Afzal, Z., van Mulligen, E., and Kors, J. (2015). Religator: chemical-disease relation extraction using prior knowledge and textual information. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Sevilla, Spain*, pages 247–253.

Przybyła, P., Shardlow, M., Aubin, S., Bossy, R., Eckart de Castilho, R., Piperidis, S., McNaught, J., and Ananiadou, S. (2016). Text mining resources for the life sciences. *Database*, 2016(0):baw145.

Quan, C., Wang, M., and Ren, F. (2014). An unsupervised text mining method for relation extraction from biomedical literature. *PloS one*, 9(7):e102039.

Wang, X., Yang, C., and Guan, R. (2015). A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics*, Sep.

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., and Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *bioRxiv*.