# Analysing Entity Type Variation across Biomedical Subdomains

**Claudiu Mihăilă, Riza Theresa Batista-Navarro, Sophia Ananiadou**

National Centre for Text Mining
School of Computer Science, University of Manchester
Manchester Interdisciplinary Biocentre,
131 Princess Street, M1 7DN, Manchester, UK
Email: {claudiu.mihaila, riza.batista-navarro}@cs.man.ac.uk,
sophia.ananiadou@manchester.ac.uk

**Abstract**

Previous studies have shown that various biomedical subdomains have lexical, syntactic, semantic and discourse structure variations. It is essential to recognise such differences to understand that biomedical natural language processing tools, such as named entity recognisers, that work well on some subdomains may not work as well on others. In this paper, we investigate the pairwise similarity (or dissimilarity) amongst twenty selected biomedical subdomains, at the level of named entity types. We evaluate the contribution of these types in the classification task by computing the chi-squared statistic over their distributions. We then build a binary classifier for each possible pair of subdomains, the results of which indicate the subdomains that are highly different or similar to others. The findings can be of potential use to those building or using named entity recognisers in determining which types of named entities need to be taken into consideration or in adapting already existing tools.

**Keywords:** named entity, subdomain variation, machine learning, biomedical text mining

## 1. Introduction

Statements regarding associations and connections between biological events and processes are central to identifying facts and claims of interest in biomedical science. Both events and processes are created on top of biological entities, so it is necessary to recognise the latter with the highest possible precision. Thus, the development of tools and resources for the automatic analysis of named entities (NEs) is key to information extraction (IE) and text mining for domain-specific scientific text.

In the past decade, researchers have focussed on fundamental tasks needed to create intelligent systems capable of improving search engine results and easing the work of biologists. More specifically, researchers have concentrated mainly on named entity recognition, normalisation to specialised databases (Krallinger et al., 2008) and extracting simple binary relations between entities.

Whilst a multitude of tools and resources have been introduced in domain-specific natural language processing (NLP) efforts for the recognition of entity mentions in text, a high proportion of these was trained and evaluated on popular corpora such as BioInfer (Pyysalo et al., 2007), GENETAG (Tanabe et al., 2005), GENIA (Kim et al., 2008), and PennBioIE (Kulick et al., 2004), as well as shared task corpora from BioCreative I, II, III (Arighi et al., 2011) and BioNLP 2009 and 2011 (Kim et al., 2011). Most of these corpora consist of documents from the molecular biology subdomain. However, previous studies (discussed in Section 2) have established that different biomedical sublanguages exhibit linguistic variations. It follows that tools which were developed and evaluated on corpora derived from one subdomain might not always perform as well on corpora from another subdomain. Understanding these linguistic variations is essential to domain adaptation of natural language processing tools.

In this paper, we highlight the similarities and differences found between biomedical sublanguages by focussing on the various types of named entities that are relevant to them. We show that for some pairs of subdomains, the frequencies of their named entity types are very similar, implying that these subdomains are very closely related. For others, however, the frequencies of different named entity types are diverse enough to allow a classifier for biomedical subdomains to be built based upon them.

This study is performed on open access journal articles found in the UK PubMed Central (UKPMC) (McEntyre et al., 2010), an article database that extends the functionality of the original PubMed Central (PMC) repository[1]. This database was chosen as our source, as most of the documents it contains are already tagged with named entity information. Reported in this paper are results obtained for 8,000 articles from 20 different biomedical subdomains.

## 2. Related Work

The work of Harris (1968) introduced a formalisation of the notion of sublanguage, which he defined as a subset of general language. According to his theory, it is possible to process specialised languages, since they have a structure that can be expressed in a computable form. Several works on the study of biomedical languages substantiated his theory, including the work of Sager et al. (1987) on pharmacological literature and lipid metabolism, and that of Friedman et al. (2002) analysing the properties of clinical and biomolecular sublanguages.

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

Other studies have investigated the differences between general and biomedical languages by focussing on specific linguistic aspects, such as verb-argument relations and pronominal anaphora. For instance, Wattarujeekrit et al. (2004) analysed the predicate-argument structures of 30 verbs used in biomedical articles. Their results suggest that, in certain cases, a significant difference exists in the predicate frames compared to those obtained from analysing news articles in the PropBank project (Palmer et al., 2005). Similarly, based on the GENIA and PennBioIE corpora, Cohen et al. (2008) perform a study of argument realisation with respect to the nominalisation and alternation of biomedical verbs. They conclude that there is a high occurrence of these phenomena in this semantically restricted domain, and underline that this sublanguage model applies only to biomedical language.

Taking a different angle, Stetson et al. (2002) uncovered the differences between "signout" notes and other medical notes (e.g., ambulatory clinic notes and discharge summaries) in terms of three aspects: discourse length, abbreviation use and abbreviation ambiguity. Based on their findings, "signout" notes are shorter and use a higher number of less ambiguous abbreviations. Nguyen and Kim (2008), on the other hand, examined the differences in the use of pronouns in general and biomedical domains by studying the MUC, ACE and GENIA corpora. They observed that compared to the MUC and ACE corpora, the GENIA corpus has significantly more occurrences of neutral and third-person pronouns, whilst first and second person pronouns are non-existent.

Verspoor et al. (2009) measured the lexical and structural variation in biomedical Open Access journals and subscription-based journals, concluding that there are no significant differences between them. Therefore, a model trained on one of these sources can be used successfully on the other, as long as the subject is maintained. Furthermore, they compare a mouse genomics corpus with two reference corpora, one composed of newswire texts and another of general biomedical articles. In this case, unsurprisingly, significant differences are found across many linguistic dimensions. Relevant to our study is the comparison between the more specific mouse genome corpus to the more general biomedical one: whilst similar from some points of view, such as negation and passivisation, they differ in sentence length and semantic features, such as the presence of various named entities.

This study, in contrast, investigates the differences and similarities between any two of twenty biomedical sublanguages at the level of named entities. Examining the distributions of different named entity types across several categories, our work is subtly similar to that of Cohen et al. (2010) who looked at the distributional variations of semantic classes in their effort to characterise the differences between abstracts and full texts. Four semantic classes, namely, *Gene*, *Mutation*, *Drug* and *Disease*, were taken into account in their study. Except for *Gene*, significant differences in terms of densities per thousand words have been observed between abstracts and full texts.

Also relevant is the work of Lippincott et al. (2011) in which a clustering-based quantitative analysis of the linguistic variations across 38 different biomedical sublanguages was presented. They investigate four dimensions relevant to the performance of NLP systems, i.e. vocabulary, syntax, semantics and discourse structure. With regard to semantic features, the authors induced a topic model using Latent Dirichlet Analysis for each word, and then extended the model to documents and subdomains according to observed distributions. Their conclusion is that an unsupervised machine learning system is able to create robust clusters of subdomains, thus proving their hypothesis that the commonly used molecular biology subdomain is not representative of the domain as a whole. In contrast, we examine the differences and similarities between biomedical sublanguages at the level of named entities, using supervised machine learning algorithms and on a different number of subdomains.

## 3. Methodology

We initially created a corpus of documents from various biomedical subdomains, from which we then extracted named entity information automatically. The NEs were later transformed into input for machine learning algorithms, as discussed below.

### 3.1. Document Collection

A corpus was created by first searching the NLM Catalog[2] for journals which are in English and available via PubMed Central, and then narrowing down the results to those whose Broad Subject Term attributes contain only one biomedical subdomain name. Since we are interested in full-text articles, we retained only those journals which are available within the PubMed Open Access subset[3]. After obtaining the total number of documents across different journals in each subdomain, we retained only those subdomains with at least 400 documents.

Using the PMC IDs of all articles under the 20 remaining subdomains, we retrieved documents from UKPMC. For each subdomain, we selected the first 400 documents with the largest number of annotated named entities. The retrieved documents are in XML format. Several unusable fragments were removed before converting them to plain text. Examples of such fragments are article metadata (authors, affiliations, publishing history), tables, figures, and references. Table 1 shows the 20 subdomains and the approximate size of the corresponding corpus subset (in number of words) after the pre-processing step.

### 3.2. Tagging of Named Entities

We formed a silver standard corpus by harmonising the annotations of multiple resources and named entity recognisers. This method was chosen due to the fact that there are no gold standard annotations available for such a large number of full-text articles.

To create the named-entity-tagged corpus, we used a simple method that augments the named entities present in the UKPMC articles with the output of two named entity recognition tools (NERs), i.e. NeMine and OSCAR. In UKPMC,

[2] http://www.ncbi.nlm.nih.gov/nlmcatalog

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

| Subdomain | Shortname | No. of words |
|---|---|---|
| Allergy and Immunology | Allergy | 0.9M |
| Biology | Biology | 3.3M |
| Cell Biology | CellBio | 3.2M |
| Communicable Diseases | Communi | 1.4M |
| Critical Care | Critica | 1.6M |
| Environmental Health | Environ | 1.9M |
| Genetics | Genetic | 3.0M |
| Health Services Research | HealthS | 1.7M |
| Medical Informatics | Medical | 2.6M |
| Medicine | Medicin | 2.1M |
| Microbiology | Microbi | 2.6M |
| Neoplasms | Neoplas | 2.2M |
| Neurology | Neurolo | 2.3M |
| Pharmacology | Pharmac | 1.8M |
| Physiology | Physiol | 3.5M |
| Public Health | PublicH | 1.7M |
| Pulmonary Medicine | Pulmona | 1.9M |
| Rheumatology | Rheumat | 1.9M |
| Tropical Medicine | Tropica | 1.7M |
| Virology | Virolog | 2.3M |

Table 1: The 20 subdomains in the corpus, their shortnames and number of words in the corpus subset.

| Type | UKPMC | NeMine | OSCAR |
|---|---|---|---|
| Gene | ✓ | ✓ | |
| Protein | ✓ | ✓ | |
| Gene\|Protein | ✓ | | |
| Disease | ✓ | ✓ | |
| Drug | ✓ | ✓ | |
| Metabolite | ✓ | ✓ | |
| Bacteria | | ✓ | |
| Diagnostic process | | ✓ | |
| General phenomenon | | ✓ | |
| Indicator | | ✓ | |
| Natural phenomenon | | ✓ | |
| Organ | | ✓ | |
| Pathologic function | | ✓ | |
| Symptom | | ✓ | |
| Therapeutic process | | ✓ | |
| Chemical molecule | | | ✓ |
| Chemical adjective | | | ✓ |
| Enzyme | | | ✓ |
| Reaction | | | ✓ |

Table 2: Named entity types and their source.

only six named entity types are annotated; with the use of NeMine and OSCAR, however, we obtained a total of 19 different classes of entities, summarised in Table 2.

Named entities in the UKPMC database were identified using NeMine (Sasaki et al., 2008), a dictionary-based statistical named entity recognition system. This system was later extended and used by Nobata et al. (2009) to include more types, such as phenomena, processes, organs and symptoms. We used this most recent version of the software as our second source of more diverse entity types.

The Open-Source Chemistry Analysis Routines (OSCAR) software (Corbett and Copestake, 2008; Jessop et al., 2011) is a toolkit for the recognition of named entities and data in chemistry publications. Currently in its fourth version, it uses three types of chemical entity recognisers, namely regular expressions, patterns and Maximum Entropy Markov models.

Nevertheless, due to the combination of several NERs, some NE types are more general and comprise other more specific types, therefore leading to double annotation. For instance, the *Gene|Protein* type is more general than both *Gene* and *Protein*, so only *Gene* or *Protein* will be kept in case they overlap with *Gene|Protein*. The same applies to the *Chemical molecule* type, which is a hypernym of *Gene*, *Protein*, *Drug* and *Metabolite*. In the case of multiple annotations over the same span of text, we removed the more general *Chemical molecule* type, so that each entity is labelled only with the more specific category assigned. Although this type of multiple annotations was frequent, we did not encounter any case of contradicting annotations over the same span of text.

This corpus is available upon request from the authors.

### 3.3. Experimental Setup

Based on the corpus previously described, we created a data set for supervised machine learning algorithms. Every document in the corpus was transformed into a vector consisting of 19 features. Each of these features corresponds to an entity type in Table 2, having a numeric value ranging from 0 to 1. This value represents the ratio of the specific entity type to the total number of named entities recognised in that document, as shown in Equation 1.

$$\theta = \frac{n_{type}}{N} \quad (1)$$

, where $n_{type}$ represents the number of named entites of a certain type in a document and $N$ represents the total number of named entities in that document. Each vector was labelled with the name of the subdomain to which the respective document belongs.

From the twenty subdomains in the corpus, we formed all possible combinations of two (thus resulting in a total of 190 pairs) for each of which we built a binary classifier. Weka (Witten and Frank, 2005; Hall et al., 2009) was employed as the machine learning framework, due to its large variety of classification algorithms. We experimented with a large number of classifiers, including J48, JRip, Logistic, RandomTree, RandomForest, SMO and combinations of these with AdaBoost. Evaluation was performed using the 10-fold cross-validation technique. RandomForest obtained the best F-score in 86 out of the 190 subdomain pairs, whilst the best result in 98 cases was obtained by AdaBoost in combination with other algorithms (JRip, RandomTree, Logistic). The remaining pairs were best classified by JRip (4 pairs) and Logistic (2 pairs). We therefore decided to present in this paper only the results using RandomForest.

## 4. Results and Analysis

We initially evaluated the value of the selected features for our task with a statistical significance test, and then performed the machine-learning experiments. Finally, we discuss the obtained results.

### 4.1. Feature Evaluation

To confirm the value of the selected features in classifying documents into subdomains, we performed the chi-squared ($\chi^2$) test of independence between each named entity and each pair of subdomains. Chi-squared is defined in Equation 2, whilst the expected value of the observation is computed according to Equation 3.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \quad (2)$$

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N} \quad (3)$$

The values are obtained by applying the ChiSquare Attribute Evaluator that is implemented in Weka. Each result contains a vector of 19 chi-squared scores, one for each feature. To visualise this graphically, we computed the Frobenius norm of the vector of chi-squared values for each subdomain pair. The Frobenius norm is defined as the square root of the sum of the absolute squares of its elements, as seen in Equation 4 (Golub and van Van Loan, 1996).

$$\|A\|_F = \sqrt{AA^*} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \quad (4)$$

, where $A^*$ denotes the conjugate transpose of $A$.

The resulting heatmap is included as Figure 1. The higher the value of the Frobenius norm, the better is the combination of features for distinguishing between the two subdomains in the pair.

To gain an insight into which features contribute most or least to the overall task, the sum of the chi-squared statistic for each feature was taken over all pairs of subdomains. We present the maximum and minimum values obtained from this exercise in Table 3.

### 4.2. Classifier Results

From the 20 subdomains, a binary classifier was built for each possible subdomain pair, as discussed in the previous section. The heatmap in Figure 2 shows the performance of each of the 190 pairs in terms of F-score. This heatmap is non-symmetric, in the sense that the F-score of subdomains A and B is different from that of B and A. All F-scores presented in this heatmap are computed with respect to the subdomain on the Y-axis (left) and against the subdomains on the X-axis (top).

A cell with a dark shade of grey corresponds to a pair of subdomains which are discernible from each other by a classifier trained on named entity type frequencies. *Cell Biology* and *Pharmacology*, for example, are found to have very distinct named entity type frequencies, as evidenced by the very good performance (97.15% F-score) of the classifier for them.

| Type | Mean |
|---|---|
| Bacteria | 10.57 |
| Chemical adjective | 19.07 |
| Chemical molecule | 87.84 |
| Diagnostic process | 24.30 |
| Disease | 195.06 |
| Drug | 82.57 |
| Enzyme | 30.77 |
| Gene | 78.03 |
| Gene\|Protein | 145.94 |
| General phenomenon | 0.34 |
| Indicator | 63.10 |
| Metabolite | 112.17 |
| Natural phenomenon | 7.07 |
| Organ | 35.78 |
| Pathologic function | 5.79 |
| Protein | 140.83 |
| Reaction | 108.43 |
| Symptom | 16.46 |
| Therapeutic | 56.09 |

Table 3: Mean values of the chi-squared statistic for each feature over all pairs of subdomains.

On the other hand, a lighter tint of grey means that the corresponding pair consists of subdomains which are very similar in their named entity type frequencies. Such is true in the case of *Communicable Diseases* and *Tropical Diseases*, for instance, in which the classifier obtained an F-score of 56.63%.

### 4.3. Analysis

From these results, we are able to enumerate the subdomains which can be considered as different or similar to a subdomain of interest in terms of frequencies of their named entity types. In obtaining the most similar subdomains, we looked at the pairs whose F-score is at the lower end of the scale. There are no pairs for which the F-scores are between 50 to 55%, and only two pairs fall within the 55-60%-range. We hence used as threshold an F-score of 65% (i.e., subdomains in pairs for which the F-score of the classifier is 65% and below were considered similar). On the other hand, we looked at the other end of the scale (i.e., pairs for which the F-score of the classifier is 95% and above) to obtain a listing of the most dissimilar subdomains.

Findings in Table 4 suggest that when building NLP tools (e.g., named entity recognisers) for documents under the subdomain in the first column, one might trivially adapt those developed for the corresponding subdomains in the second column. A named entity recogniser for the *Microbiology* subdomain, for example, might be trivially applied to *Neoplasms* documents. However, it might also be the case that there are no named entity recognisers built yet that are specialised for these subdomains.

In contrast, those built for the subdomains in the second column of Table 5 might need further training or adaptation in applying them to the corresponding subdomain in the first column, as these tools might have been trained on

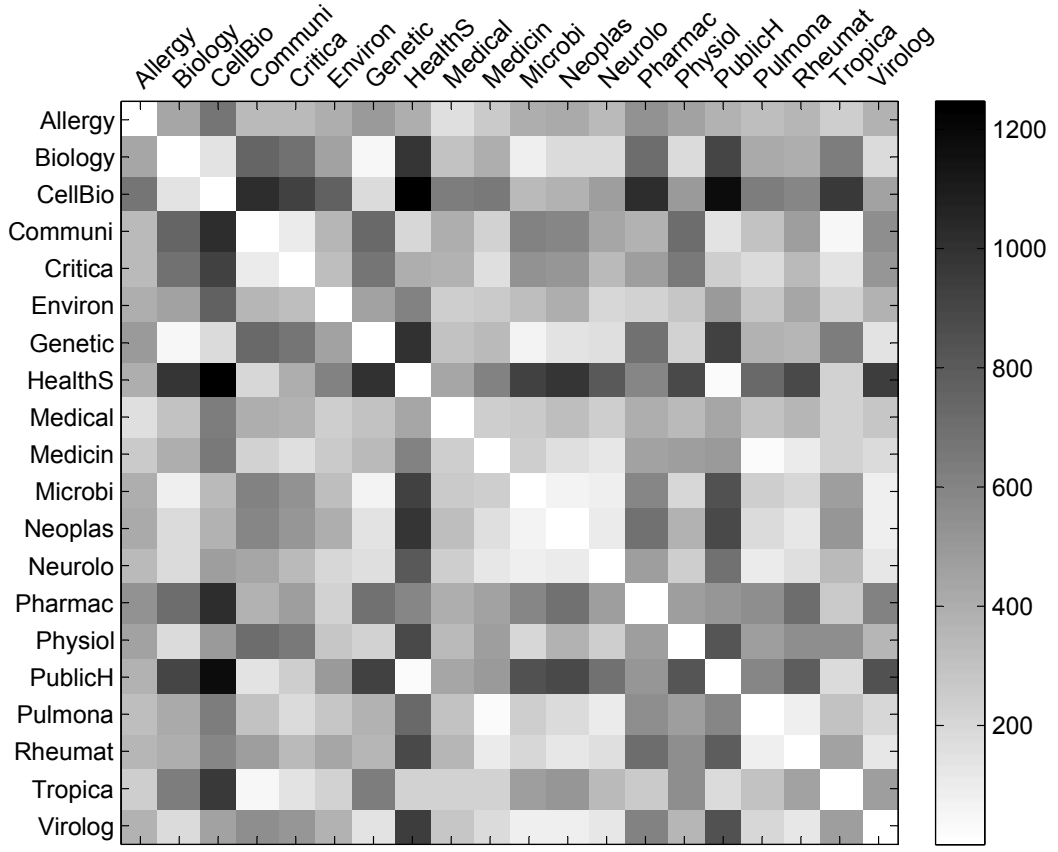


Figure 1: A heatmap showing the Frobenius norm based on the chi-squared vector for each pair of subdomains.

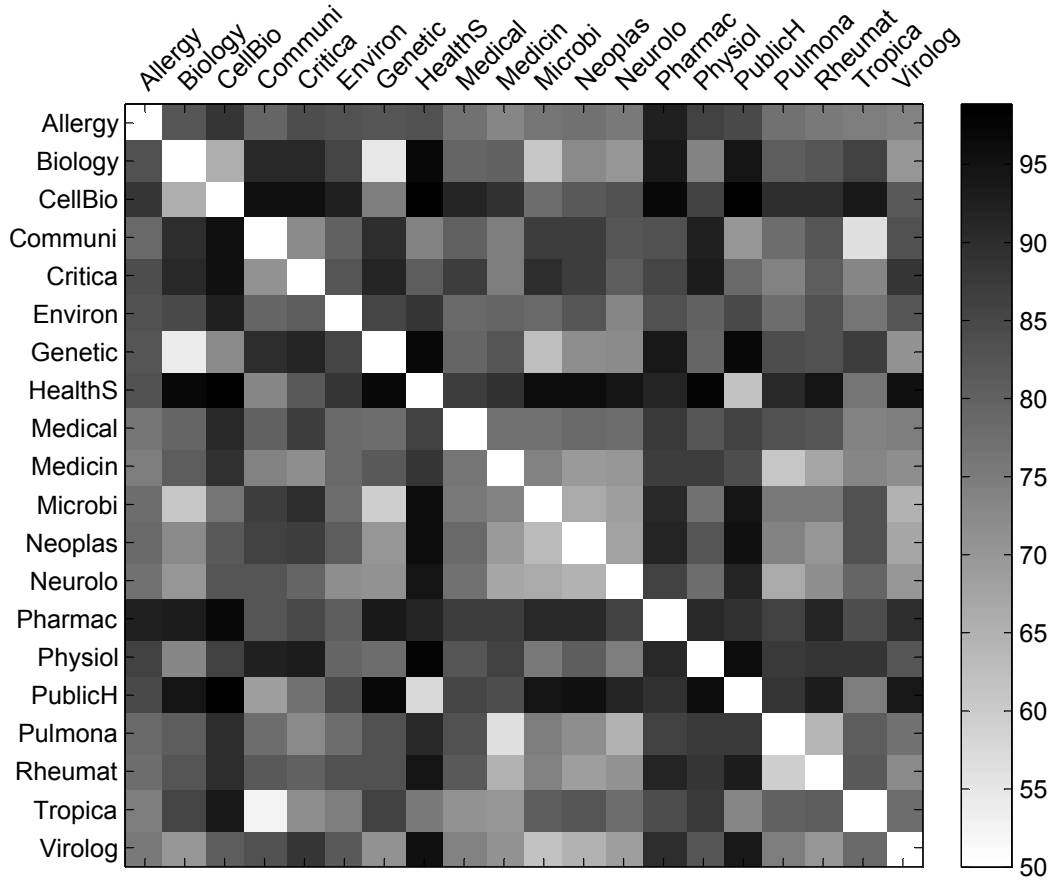


Figure 2: A heatmap showing the performance (in F-score) of each classifier built for each pair of subdomains.

| Subdomain | Similar subdomains |
|---|---|
| Biology | Cell Biology, Genetics, Microbiology |
| Communicable Diseases | Tropical Diseases |
| Medicine | Pulmonary Medicine |
| Health Services Research | Public Health |
| Genetics | Microbiology |
| Pulmonary Medicine | Rheumatology |
| Microbiology | Virology |

Table 4: Similar subdomains. The subdomains listed in the second column can be considered as highly similar to the corresponding subdomain in the first column based on their named entity type frequencies.

| Subdomain | Dissimilar subdomains |
|---|---|
| Biology | Public Health, Health Services Research |
| Cell Biology | Critical Care, Communicable Diseases, Pharmacology, Public Health, Health Services Research |
| Genetics | Public Health, Health Services Research |
| Health Services Research | Microbiology, Neoplasms, Physiology, Rheumatology, Virology |
| Neoplasms | Public Health |
| Physiology | Public Health |

Table 5: Dissimilar subdomains. The subdomains listed in the second column can be considered as different from the corresponding subdomain in the first column based on their named entity type frequencies.

documents where the named entity types which occur frequently in the subdomain of interest, are sparse. For instance, there is no certainty that NERs developed for the *Pharmacology* domain will work well on *Neoplasms* documents.

We computed the mean along each row and column of the heatmap, and determined that both the row and column corresponding to *Medicine* produced the minimum, while *Pharmacology* has the maximum. This finding suggests that *Medicine* is the biomedical subdomain which is most "alike" every other subdomain, irrespective of the direction F-score is computed in, while *Pharmacology* is the least one. In developing a named entity recogniser for *Pharmacology*, one has to consider its differences with other biomedical subdomains in terms of named entity type distributions.

## 5. Conclusion

We formed a silver standard corpus from 20 biomedical subdomains and built a binary classifier for each possible subdomain pair. From the results, we have observed which subdomains are highly discernible from each other by a classifier, in terms of named entity type frequencies. However, there are also cases when a classifier is unable to distinguish between subdomains, implying that they have highly similar named entity type distributions.

Such differences and similarities in named entity type frequencies should be considered when developing automated tools for one subdomain and adapting them for use on another.

## 6. References


Cecilia Arighi, Zhiyong Lu, Martin Krallinger, Kevin Cohen, W Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy Wu. 2011. Overview of the BioCreative III Workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Kevin Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9):e3158, 09.

Kevin Bretonnel Cohen, Helen Johnson, Karin Verspoor, Christophe Roeder, and Lawrence Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

Peter Corbett and Ann Copestake. 2008. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, 9(Suppl 11):S4.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Gene H. Golub and Charles F. van Van Loan. 1996. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences) (3rd Edition)*. The Johns Hopkins University Press, 3rd edition, October.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Zellig Harris. 1968. *Mathematical Structures of Language*. John Wiley and Son, New York.

David Jessop, Sam Adams, Egon Willighagen, Lezan Hawizy, and Peter Murray-Rust. 2011. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):41.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA, June. Association for Computational Linguistics.

Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. 2008. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology*, 9(Suppl 2):S1.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of the BioLINK 2004*.

Thomas Lippincott, Diarmuid Seaghdha, and Anna Korhonen. 2011. Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, 12(1):212.

Johanna R. McEntyre, Sophia Ananiadou, Stephen Andrews, William J. Black, Richard Boulderstone, Paula Buttery, David Chaplin, Sandeepreddy Chevuru, Norman Cobley, Lee-Ann Coleman, Paul Davey, Bharti Gupta, Lesley Haji-Gholam, Craig Hawkins, Alan Horne, Simon J. Hubbard, Jee-Hyub Kim, Ian Lewin, Vic Lyte, Ross MacIntyre, Sami Mansoor, Linda Mason, John McNaught, Elizabeth Newbold, Chikashi Nobata, Ernest Ong, Sharmila Pillai, Dietrich Rebholz-Schuhmann, Heather Rosie, Rob Rowbotham, C. J. Rupp, Peter Stoehr, and Philip Vaughan. 2010. UKPMC: a full text article resource for the life sciences. *Nucleic Acids Research*.

Ngan L. T. Nguyen and Jin-Dong Kim. 2008. Exploring domain differences for the design of pronoun resolution systems for biomedical text. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 625–632, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chikashi Nobata, Yutaka Sasaki, Naoaki Okazaki, C.J. Rupp, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Semantic search on digital document repositories based on text mining results. In *International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)*, pages 34–48.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50.

Naomi Sager, Carol Friedman, and Margaret Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data*. Addison-Wesley, Reading, MA.

Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou. 2008. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9(Suppl 11):S5.

Peter D Stetson, Stephen B Johnson, Matthew Scotch, and George Hripcsak. 2002. The sublanguage of cross-coverage. *Proceedings of the AMIA Symposium*, pages 742–746.

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

Karin Verspoor, Kevin Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, 10(1):183.

Tuangthong Wattarujeekrit, Parantu Shah, and Nigel Collier. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(1):155.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.