A dictionary- and rule-based system for identification of bacteria and habitats in text

Helen V Cook
Novo Nordisk Foundation
Center for Protein Research,
Faculty of Health and Medical Sciences,
University of Copenhagen, Denmark
helen.cook@cpr.ku.dk

Evangelos Pafilis
Institute of Marine Biology,
Biotechnology and Aquaculture,
Hellenic Centre for Marine Research,
Crete, Greece
pafilis@hcmr.gr

Lars Juhl Jensen
Novo Nordisk Foundation
Center for Protein Research,
Faculty of Health and Medical Sciences,
University of Copenhagen, Denmark
lars.juhl.jensen@cpr.ku.dk

Abstract

The number of scientific papers published each year is growing exponentially and given the rate of this growth, automated information extraction is needed to efficiently extract information from this corpus. A critical first step in this process is to accurately recognize the names of entities in text. Previous efforts, such as SPECIES, have identified bacteria strain names, among other taxonomic groups, but have been limited to those names present in NCBI taxonomy. We have implemented a dictionary-based named entity tagger, TagIt, that is followed by a rule based expansion system to identify bacteria strain names and habitats and resolve them to the closest match possible in the NCBI taxonomy and the OntoBiotope ontology respectively. The rule based post processing steps expand acronyms, and extend strain names according to a set of rules, which captures additional aliases and strains that are not present in the dictionary. TagIt has the best performance out of three entries to BioNLP-ST BB3 cat+ner, with an overall SER of 0.628 on the independent test set.

1 Introduction

The biomedical literature is growing at an estimated 4% per year and as of 2016 there are at least 26 Million documents in PubMed (Lu, 2011). 12% of this work is never cited after 5 years and much of it might not reach its intended audience, effectively limiting the value of these scientific contributions (Lariviere et al., 2008). Molecular biology databases such as UniProt address this issue by manually curating domain-specific knowledge and providing it in a structured form (The UniProt Consortium, 2014). Despite efforts by the metagenomics community (Lombardot et al., 2006; Reddy et al., 2015; Hoopen et al., 2016), the same attention has not been given to manual curation in microbial and molecular ecology, where a lack of samples annotated with comprehensive metadata hinders comparative and integrative studies (Yilmaz et al., 2011). Both the initial creation and subsequent ongoing maintenance of such databases require a significant investment of labour and money (Attwood et al., 2015). In order to scale up this process, we need to automate the extraction of information from text.

The BioCreative and BioNLP communities are responding to this need by organising scientific literature mining challenges that aim to advance the state of the art (Arighi et al., 2014; Bossy et al., 2015). These competitions have resulted in the development of text-mining tools focusing on specific curation tasks (Bossy et al., 2015; Wang et al., 2015), one of which is the interactive EXTRACT tool that assists curators through automated named entity recognition (NER) of organisms, tissues, diseases and environments (Pafilis et al., 2015).

The BioNLP BB3 focuses on the identification of bacteria and their habitats in text. Bacteria are ubiquitous in natural and artificial environments, and play major diverse roles in shaping ecosystems. They thrive in the most extreme habitats – under the west Antarctic ice sheet (Christner et al., 2014), in alkaline hot springs (De León et al., 2013) – and they also proliferate in the most mundane habitats – such as the human body, which contains roughly an equal number of bacterial and human cells (Sender et al., 2016). Bacteria are responsible for the majority of nitrogen fixation on the planet (Galloway et al., 2004), affect the absorption of nutrients in the human gut (Semova et al., 2012), and are responsible for the

Proceedings of the 4th BioNLP Shared Task Workshop, pages 50–55,
Berlin, Germany, August 13, 2016.

deaths of approximately 1.5 million people each year from *Mycobacterium tuberculosis* infection (WHO, 2016). Given both their beneficial and pestilential impacts, it is important to understand the habitats in which bacteria grow so that they can be managed and controlled, especially in medical environments that provide care for immunocompromised patients (Sydnor and Perl, 2011), and in food processing environments which have the potential for wide distribution of contaminated products (Brackett, 1999).

The first steps towards automatically turning unstructured text into structured information about bacteria and their habitats are i) to recognize names of bacteria and habitats in a text, and ii) to resolve these to a predefined ontology or taxonomic resource. Whereas the first step can be addressed in a variety of different ways, such as using machine learning, manually crafted rules or dictionaries, the second step clearly requires the use of a dictionary.

The SPECIES and ORGANISMS resources are purely dictionary based methods that demonstrate above 85% precision and recall on identifying cellular organisms in abstracts (Pafilis et al., 2013). Further, these tools have extremely fast run times, a necessary requirement for processing large datasets. Dictionary based methods have the advantage of always correctly normalizing a term that has been tagged, but conversely they have the disadvantage of requiring an up-to-date, comprehensive dictionary. Building such a dictionary can be a difficult manual task, but it can be aided by the use of orthographic expansion rules and stopword lists. When parsing documents from a limited domain, such as biomedical literature, the dictionary required is much smaller in scope, and building one becomes feasible, as has been demonstrated by SPECIES and ORGANISMS which have been built from NCBI Taxonomy (Sayers et al., 2009).

NCBI taxonomy is a curated classification and nomenclature resource that covers all of the organisms in the Entrez sequence database (Sayers et al., 2009). Although these resources are the most comprehensive of their kind, very new and very old strains that are lacking sequences cannot be found in the NCBI taxonomy, and neither can known strains that have been spelled with uncommon misspellings. Further, acronyms that are not defined as synonyms will also not be present, meaning that a dictionary method that naively used only the entries in the taxonomy would miss tagging such terms.

Here we present TagIt, a tool for named entity recognition and categorization of bacteria and their habitats. It primarily uses a dictionary-based approach, the results of which are extended with pattern-matching rules that handle acronyms that are not found in the dictionary and refine match boundaries to include bacterial strain names.

2 Methods

2.1 Dictionary creation

A dictionary for bacteria terms was generated from all NCBI taxonomy entities under the bacteria superkingdom (taxid: 2) (Sayers et al., 2009). The dictionary generation process is based on that used in (Pafilis et al., 2013). Briefly, NCBI taxonomy provides alternate names for each taxonomy level, which include common names, obsolete names and other synonyms, all of which were included in our dictionary. These terms were expanded to plural forms following the English and Latin rules for pluralizing nouns, and the abbreviations of Linnaean names, such as *E. coli* for *Escherichia coli*, were generated and included in the dictionary.

A dictionary for habitat terms was generated from the OntoBiotope ontology (OBT), and the names present in the ontology were expanded to their plural forms giving 8,345 terms. The habitat dictionary was expanded via synonym transfer based on manual mappings between OBT terms and their Brenda Tissue Ontology (BTO) counterparts (Chang et al., 2015). The BTO name dictionary available in the TISSUES database facilitated this process (Santos et al., 2015). This gave an additional 121,321 habitat synonyms. For example, the term "central nervous system" (OBT:000831) was expanded to include "hippocampus" and 2748 other terms, 76 of which are particular cell lines derived from nervous system tissue.

The same synonym transfer process was applied to map OBT terms to their NCBI taxonomy counterparts under the eukaryote branch (taxid: 2759). The term duck (OBT:002200), for example, was expanded with 46 synonyms including "mallard ducks", "northern mallard", "*Anas platyrhynchos*", and so forth. Terms that existed in NCBI taxonomy but not in OBT were mapped to the most specific relevant term. For example, all 145,546 names and synonyms under the NCBI

taxonomy node Metazoa (taxid: 33208) that could not be mapped to anything more specific in OntoBiotope were mapped to "animal" (OBT:000218). This gave a total of 5,106,213 additional synonyms.

Synonym transfer was also applied to OBT and the corresponding Environments Ontology (ENVO) terms (with name information from the ENVIRONMENTS tool) for an additional 54,673 synonyms (Buttigieg et al., 2013; Pafilis et al., 2015). However, as shown later, this did not improve the systems accuracy and so was not used in the final version.

Since dictionary-based NER is prone to poor precision, especially after automatic dictionary expansion, stopword lists are used to remove matches that contribute the most to the drop in precision. Here, stopword lists were generated for both bacteria and habitat entity types by manually inspecting the most frequently identified terms when tagging the Medline corpus, and removing those terms that were likely to not refer to true positive matches. This resulted in 2381 stopwords for bacteria including words such as "unclassified", and 2592 stopwords for habitat, including words such as "scales" and "root", which can have many different meanings. The full dictionaries, including the stopwords, are provided in the associated repository located at http://github.com/bitmask/BioNLP-BB3.

2.2 Tagging and post processing

Both entity types were tagged using the left-most longest matching and hashing function present in the SPECIES tool, which is case insensitive, and disregards hyphens and white space characters within names and quotes and parentheses at the beginning or end of names (Pafilis et al., 2013).

A series of post processing steps followed the tagging step. First, the input document was examined for parentheses, and these and their contents were replaced by whitespace. The tagger was run again on the modified text to identify any additional matches that spanned the parentheses. These new results were merged into the original results.

Second, the normalizations were filtered to return only the highest confidence normalization for each entry (by default SPECIES may return multiple normalizations). The normalizations for bacteria were updated so that a mention of a genus that

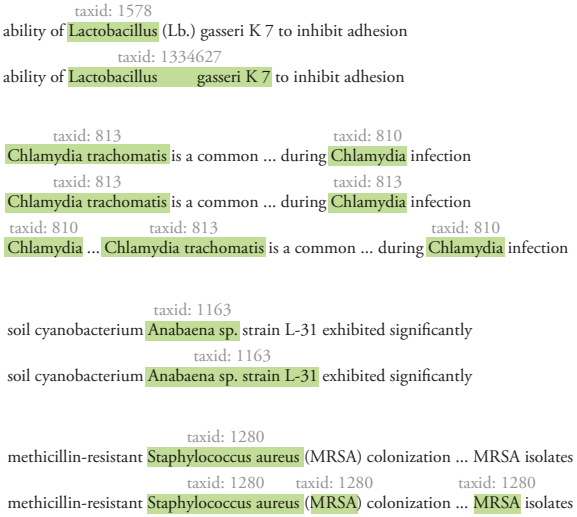

Figure 1: Illustration of the four post processing steps: Parentheses avoidance, normalization correction, strain expansion, and acronym expansion, where the first line in each block indicates the matches and normalizations prior to post processing, and the subsequent lines show how they are updated after post processing.

followed a more specific species mention (within that genus) would be normalized to the species. Although not in the annotation guidelines, we added the exception that if the genus was mentioned alone before any species within that genus, then later mentions of the genus would not be changed to refer to the specific species because such mentions were much more likely to refer to the genus in general than to have been an instance of synecdoche. These cases are illustrated in Figure 1.

Third, for bacteria, strain names were expanded by matching the text immediately following a match returned from the tagger against a regex that would identify it as a strain. Strains names were identified as sequences of letters and punctuation that may have included an indicator such as "sp." or "strain".

Lastly, acronyms were identified for both bacteria and habitats by searching the text following a match for a potential short form. Text was considered to be a short form if it was within parentheses, contained capital letters, and contained the first letter of the long form within its first three letters. Then, the remainder of the document was searched for further instances of the short form, which were normalized to the definition of the long form.

Full details and code are available at http://github.com/bitmask/BioNLP-BB3

3 Results and Discussion

Our entry, TagIt, performed best out of three entries submitted to the BioNLP-ST BB3-cat+ner task with an overall slot error rate (SER) of 0.628 on the test set. For bacteria only the SER was 0.399, and for habitats only the SER was 0.775.

TagIt uses a dictionary for both named entity recognition and for categorization, which is generated a priori from existing ontologies and rules regarding name expansion. Generating the dictionary does not require the input of any training documents, nor does this approach require that the values of any variables be learned during a training step. Therefore, we have evaluated our method on both the provided training and development sets, and see consistent performance between them.

In order to quantify the improvements from expanding the dictionaries, we generated six iterations of the dictionary that we evaluated independently on the training and development sets. The first, included only the dictionary for bacteria. The second naively added in habitats from the OntoBiotope ontology with no synonym transfer for the habitats dictionary. The next three variants transferred synonyms to the habitats dictionary from BTO, eukaryotic entries from NCBI and ENVO, respectively. The final dictionary featured synonym transfer from both BTO and NCBI, giving better performance than either alone. This dictionary was selected as our final submission to the contest.

For both training and development sets, performance increased (i.e. SER decreased) with the addition of BTO and NCBI synonyms to the dictionary. The improvement in habitat only SER from using an unexpanded habitats dictionary, and including the mappings from BTO and NCBI – from 0.568 to 0.511 (dev) or 0.635 to 0.587 (train) – shows the performance increase possible by using synonyms in other ontologies to expand the range and number of synonyms present in the dictionary.

Adding ENVO synonyms surprisingly did not increase performance. The performance of this dictionary was evaluated in (Pafilis et al., 2015) at 87.8% precision over 600 documents, so it is unlikely that the lack of performance increase we see is due to some underlying defect in the dictionary. Further, the mapping between OBT and ENVO orthologies was performed manually by subject matter experts, so this is also unlikely to be a major source of error. The addition of ENVO synonyms cannot increase the false negative rate, as adding names to the dictionary will not result in less being found. The addition of ENVO synonyms did increase the false positive rate. The false positives included three terms that were used as homonyms such as "reservoir", intended in the dictionary to refer to a body of water, but used in the text to mean a source of bacteria. The instances of these false positives could be reduced by adding these terms to the stopword list. One further case registers as a false positive ("farms" at position 502, 507 in `BB-cat+ner-2696427.txt`), but upon manual inspection appears to be consistent with the annotation guidelines. Overall, the addition of the ENVO dictionary resulted in the identification of only a few additional terms, and if the identified errors were fixed, we would see only a minor improvement in performance compared to a dictionary without ENVO included.

In terms of the results for bacteria, the false negatives identified by TagIt included the names of strain mutants (such as Ara+), multiposition matches, and acronyms that are defined in a non-standard manner. Bacterial false positives included a small number of cases in which terms such as "cyanobacterium" were used as adjectives or descriptions and should not have been annotated. Further, TagIt identified an additional 3 instances in which the boundaries disagreed with the gold standard, and 27 cases in which the normalizations disagreed with the gold standard, but in both cases our annotations more closely reflected the annotation guidelines.

4 Conclusions

Accurate identification of entities in text is a first necessary step towards automated extraction of information about those entities. Here, we have presented a dictionary- and rule-based system, called TagIt, to identify bacterial names and habitats which gives good performance on both entity types.

Dictionary methods for named entity recognition and categorization can give very good performance on limited domains, and rule based post processing can help overcome the intrinsic limitations to the dictionary approach. To recognize bacterial entities, applying simple rules to expand

	Overall SER			Bacteria only SER			Habitats only SER		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
Bacteria	0.778	0.757		0.341	0.303		n/a	n/a	
Bacteria + Habitats	0.537	0.477		0.341	0.303		0.635	0.568	
Bacteria + Habitats + BTO	0.529	0.468		0.341	0.303		0.623	0.555	
Bacteria + Habitats + NCBI	0.514	0.448		0.341	0.303		0.599	0.524	
Bacteria + Habitats + ENVO	0.540	0.479		0.341	0.303		0.639	0.572	
Bacteria + Habitats + BTO + NCBI	0.506	0.439	0.628	0.341	0.303	0.399	0.587	0.511	0.775

Table 1: Performance of TagIt in terms of overall, bacteria only and habitat only slot error rates for training, development and test sets over six variations of the dictionary (see text for their definitions).

strains and acronyms helped identify names that were not present in the dictionary. Dictionary synonym expansion also increases the performance of dictionary based methods, as was seen by the addition of BTO and NCBI synonyms to our habitats dictionary, boosting the performance over what was possible with no synonym expansion.

Acknowledgments

EU BON (EU FP7 Contract No. 308454 program), the Micro B3 Project (287589), the Earth System Science and Environmental Management COST Action (ES1103) and the Novo Nordisk Foundation (NNF14CC0001).

References

Cecilia N Arighi, Cathy H Wu, Kevin B Cohen, Lynette Hirschman, Martin Krallinger, Alfonso Valencia, Zhiyong Lu, John W Wilbur, and Thomas C Wiegers. 2014. BioCreative-IV Virtual Issue. *Database*, 2014:1–6.

Teresa Attwood, Bora Agit, and Lynda Ellis. 2015. Longevity of Biological Databases. *EMBnet.journal*, 21(0).

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the Gene Regulation Network and the Bacteria Biotope Tasks in BioNLP'13 Shared Task. *BMC Bioinformatics*, 16(Suppl 10):S1.

Robert E. Brackett. 1999. Incidence, Contributing Factors, and Control of Bacterial Pathogens in Produce. *Postharvest Biology and Technology*, 15(3):305–311.

Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, and Suzanna E Lewis. 2013. The Environment Ontology: Contextualising Biological and Biomedical Entities. *Journal of Biomedical Semantics*, 4(43).

Antje Chang, Ida Schomburg, Sandra Placzek, Lisa Jeske, Marcus Ulbrich, Mei Xiao, Christoph W. Sensen, and Dietmar Schomburg. 2015. BRENDA in 2015: Exciting developments in its 25th Year of Existence. *Nucleic Acids Research*, 43(D1):D439–D446.

Brent C. Christner, John C. Priscu, Amanda M. Achberger, Carlo Barbante, Sasha P. Carter, Knut Christianson, Alexander B. Michaud, Jill A. Mikucki, Andrew C. Mitchell, Mark L. Skidmore, Trista J. Vick-Majors, and the WISSARD Science Team. 2014. A Microbial Ecosystem Beneath the West Antarctic Ice Sheet. *Nature*, 512(7514):310–313.

Kara Bowen De León, Robin Gerlach, Brent M. Peyton, and Matthew W. Fields. 2013. Archaeal and Bacterial Communities in Three Alkaline Hot Springs in Heart Lake Geyser Basin, Yellowstone National Park. *Frontiers in Microbiology*, 4(330):1–10.

J. N. Galloway, F. J. Dentener, D. G. Capone, E. W. Boyer, R. W. Howarth, S. P. Seitzinger, G. P. Asner, C. C. Cleveland, P. A. Green, E. A. Holland, D. M. Karl, A. F. Michaels, J. H. Porter, A. R. Townsend, and C. J. Vörösmarty. 2004. *Nitrogen Cycles: Past, Present, and Future*, volume 70. Kluwer Academic Publishers.

Ten Hoopen, Amid C, Luigi Buttigieg, Pafilis E, Bravakos P, Cerdeño-Tárraga AM, Gibson R, Kahlke T, Legaki A, Narayana Murthy, Papastefanou G, Pereira E, Rossello M, Luisa Toribio, and Cochrane G. 2016. Value, but High Costs in Post-Deposition Data Curation. *Database*, pages 1–10.

Vincent Lariviere, Yves Gingras, and Eric Archambault. 2008. The Decline in the Concentration of Citations, 1900-2007. *Pre-print*, (arXiv:0809.5250 [physics.soc-ph]):1–9.

Thierry Lombardot, Renzo Kottmann, Hauke Pfeffer, Michael Richter, Hanno Teeling, Christian Quast, and Frank Oliver Glöckner. 2006. Megx.net–Database Resources for Marine Ecological Genomics. *Nucleic Acids Research*, 34:D390–D393.

Zhiyong Lu. 2011. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database*, 2011:1–13.

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini

Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS ONE*, 8(6):2–7.

Evangelos Pafilis, Sune P Frankild, Julia Schnetzer, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Katerina Vasileiadou, Patrick Leary, Jennifer Hammock, Katja Schulz, Cynthia Sims Parr, Christos Arvanitidis, and Lars Juhl Jensen. 2015. ENVIRONMENTS and EOL: Identification of Environment Ontology Terms in Text and the Annotation of the Encyclopedia of Life. *Bioinformatics*, 31(11):1872–1874.

Tatiparthi B. K. Reddy, Alex D. Thomas, Dimitri Stamatis, Jon Bertsch, Michelle Isbandi, Jakob Jansson, Jyothi Mallajosyula, Ioanna Pagani, Elizabeth A. Lobos, and Nikos C. Kyrpides. 2015. The Genomes OnLine Database (GOLD) v.5: A Metadata Management System Based on a Four Level (Meta)Genome Project Classification. *Nucleic Acids Research*, 43(D1):D1099–D1106.

Alberto Santos, Kalliopi Tsafou, Christian Stolte, Sune Pletscher-Frankild, Seán I O'Donoghue, and Lars Juhl Jensen. 2015. Comprehensive Comparison of Large-Scale Tissue Expression Datasets. *PeerJ*, 3:e1054.

Eric W Sayers, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. 2009. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37:D5–15.

Ivana Semova, Juliana D. Carten, Jesse Stombaugh, Lantz C. MacKey, Rob Knight, Steven A. Farber, and John F. Rawls. 2012. Microbiota Regulate Intestinal Absorption and Metabolism of Fatty Acids in the Zebrafish. *Cell Host and Microbe*, 12(3):277–288.

Ron Sender, Shai Fuchs, and Ron Milo. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *bioRxiv pre-print*, pages 1–21.

Emily R M Sydnor and Trish M. Perl. 2011. Hospital Epidemiology and Infection Control in Acute-Care Settings. *Clinical Microbiology Reviews*, 24(1):141–173.

The UniProt Consortium. 2014. UniProt: A Hub for Protein Information. *Nucleic Acids Research*, 43(D1):D204–212.

Qinghua Wang, Shabbir Syed Abdul, Lara Almeida, Sophia Ananiadou, Yalbi Itzel Balderas-Martínez, Riza BatistaNavarro, David Campos, Lucy Chilton, Hui-Jou Chou, Gabriela Contreras, Laurel Cooper, Hong-Jie Dai, Juliane Fluck, Socorro Gama, Georgios Gkoutos, Afroza Khanam Irin, Lars Juhl Jensen, Silvia Jimenez, Toni Rose Jue, Ingrid Keseler, Sumit Madan, Sérgio Matos, Peter McQuilton, Matthew Mort, Jeyakumar Natarajan, Evangelos Pafilis, Emiliano Pereira, Shruti Rao, Fabio Rinaldi, David Salgado, Onkar Singh, Raymund Stefancsik, Chu-Hsien Su, Suresh Subramani, Hamsa Dhwani Tadepally, Loukia Tsaprouni, Nicole Vasilevsky, Xiaodong Wang, Andrew Chatr-aryamontri, Stan Laulederkind, Sherri Matis-Mitchell, Johanna McEntyre, Sandra Orchard, Sangya Pundir, Raul Rodriguez-Esteban, Kimberly Van Auken, Zhiyong Lu, Mary Schaeffer, Lynette Hirschman, and Cecilia Arighi. 2015. Overview of the Interactive Task in BioCreative V. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, page 20.

WHO. 2016. Tuberculosis.

Pelin Yilmaz, Jack A. Gilbert, Rob Knight, Linda Amaral-Zettler, Ilene Karsch-Mizrachi, Guy Cochrane, Yasukazu Nakamura, Susanna-Assunta Sansone, Frank Oliver Gloeckner, and Dawn Field. 2011. The Genomic Standards Consortium: Bringing Standards to Life for Microbial Ecology. *The ISME Journal*, pages 1565–1567.

A Supplemental Material

Code, dictionaries, and the mapping of BTO and NCBI taxonomy to OBT is available at: http://github.com/bitmask/BioNLP-BB3