Augmenting the Medical Subject Headings vocabulary with semantically rich variants to improve disease mention normalisation

Riza Batista-Navarro and Sophia Ananiadou

National Centre for Text Mining, School of Computer Science
University of Manchester
{riza.batista,sophia.ananiadou}@manchester.ac.uk

Abstract. We extended our existing methods for entity normalisation as part of our contribution to the Disease Named Entity Recognition and Normalisation subtask of the Chemical-Disease Relation (CDR) track of BioCreative V. Our newly proposed approach is based on the incorporation of semantics in two ways: (1) by adding corpus-derived semantic variants to the Medical Subject Headings (MeSH) vocabulary, and (2) through automatic translation of medical root words and affixes to potential variants. Results of the official evaluation of our methods show that the combination of both means for semantic enrichment gives us optimal performance on the disease name normalisation task, obtaining an F-score of 85.56%, with precision of 89.51% and recall of 81.94%. We have made our methods available in the form of a BioC-compliant Web service.

Key words: Disease name recognition, Normalisation, String similarity, Web services, Dictionary enrichment

1 Introduction

Relationships between chemical entities and diseases are crucial to the discovery of new drugs [5] as well as the understanding of adverse drug reactions [11]. However, the extraction of this type of relationship has not gained much attention from the biomedical text mining community compared to other types, e.g., protein-protein, gene-drug interactions [7, 20]. Aiming to foster the development of advanced methods for automatic chemical-disease relation extraction from literature, the Chemical-Disease Relation (CDR) track has been organised as part of the Fifth BioCreative Challenge Evaluation [23]. It consists of two subtasks: Disease Named Entity Recognition and Normalisation (DNER), and Chemical-induced Diseases relation extraction (CID). The first subtask called for the automatic recognition of disease mentions in PubMed abstracts, as well as the assignment of Medical Subject Heading (MeSH) identifiers [19] to these mentions. The second one, meanwhile, focussed on the extraction of mention pairs denoting the drug-induces-disease type of relation. In both subtasks, participants were asked to make their methods available as Web services complying either with the PubTator [22] or BioC [9] format.

The remainder of this paper discusses the methods we developed as part of our contribution to the DNER subtask. While the automatic identification of entity types such as genes/proteins and chemical molecules has already advanced [24, 15], the recognition and normalisation of disease mentions has remained a challenge, with only a handful of tools and resources having been developed thus far [17]. This can be attributed to the high variability with which disease mentions may appear in text, as well as the limited amount of available gold standard data [12]. In this work, we took a supervised approach for disease name recognition based on the training of conditional random fields (CRF) models [16]. For disease name normalisation, we extended existing string similarity-based methods by: (1) automatically compiling a semantically enriched version of the MeSH vocabulary, and (2) generating potential semantic variants by translating Greek or Latin medical root words and affixes.

2 Systems description and methods

The following describes our methods for disease name recognition and normalisation. We begin by providing an overview of our CRF-based approach to disease name recognition, and then proceed to outlining our strategy for normalisation of mentions against MeSH entries.

2.1 Disease name recognition

We cast named entity recognition as a sequence labelling task, in which individual tokens of the text are assigned labels according to the begin-inside-outside (BIO) encoding scheme. To facilitate the representation of our textual data in this manner, the corpora we exploited were first processed by the LingPipe sentence splitter [1] which segmented each document into sentences. These in turn were decomposed into tokens by the OSCAR4 Tokeniser [13], which were then assigned lemmatised forms as well as part-of-speech (POS) and chunk tags by the GENIA Tagger [21].

We used the NERsuite package [3], an implementation of CRFs, to train and apply models for sequence labelling. Each token was represented by a rich set of lexical, orthographic and semantic features. These include, for example: (1) two, three and four-character n-grams, (2) token, POS tag and lemma unigrams and bigrams within a window of 3, (4) presence of special characters, (5) capitalisation and (6) matches against semantically relevant dictionaries. Selected as sources of dictionary matches are the following controlled vocabularies or databases: MeSH, the Disease Ontology [14], Online Mendelian Inheritance in Man (OMIM) [4], the Comparative Toxicogenomics Database (CTD) [10] and the Unified Medical Language System (UMLS) [6].

2.2 Disease name normalisation

Several strategies for disease name normalisation were explored in this work. As a baseline method, our own reformatted version of the MeSH dictionary

Table 1: Number of unique names in each version of MeSH used. Only entries under the Diseases and Psychiatry/Psychology subtrees of MeSH were included.

Source	**Number of entries**
MeSH	53,839
MeSH+CDR	55,315
MeSH+CDR+NCBI	56,596

names/synonyms (and corresponding identifiers) was compiled, in which each entry has been transformed into a canonical representation based on the following series of steps:

1. conversion of all characters to lowercase
2. removal of stop words and punctuation
3. stemming of each remaining token
4. alphabetical re-ordering of tokens

The same transformation was performed on each disease mention occurring in text. To determine the MeSH identifier that should be assigned to the mention, the resulting canonical form is used to query our compiled dictionary to fetch the most similar strings according to the Jaro-Winker distance measure [8]. If the similarity score between the mention and a dictionary string is above a predefined threshold of 0.80, the latter is considered a candidate match. This, however, resulted in the retrieval of several candidates having the same score. We thus additionally applied the Levenshtein distance measure [8] in order to induce a more informative ranking of the candidates, based on which the candidate with the smallest distance was considered as the best matching MeSH entry. The mention in question is finally assigned the identifier attached to this entry.

Our newly proposed approach builds upon the previously described method and is based on the incorporation of more semantics. Firstly, two corpora, namely the CDR corpus [18] provided by the track organisers and the NCBI Disease Corpus, were used as sources of variants actually used in scientific literature which were added to MeSH by cross-referencing provided gold standard identifiers. Table 1 presents the resulting size of our different MeSH dictionary versions after the application of this method.

Secondly, we compiled a list of medical root words [2] and automatically combined them with affixes that are synonymous with terms pertaining to medical disorder such as “disease”, “deficiency”, “inflammation”, to generate potential variants that can be then matched against MeSH. If the score of the best matching candidate retrieved for a mention using string similarity is lower than a predefined threshold, the mention is checked for the occurrence of medical root words. The word “neuropathy”, for example, is broken down into “neuro” (nerve or nervous) and “pathy” (disease), based on which our method automatically generates “nervous disease”. When used to query our own compiled MeSH dictionary, “nervous disease” fetches “nervous system disease”, thus leading to the assignment of the correct identifier to “neuropathy”.

Table 2: Evaluation of our normalisation approaches on the CDR Test corpus

	Precision	Recall	F-score
Run 1	88.89	82.14	85.39
Run 2	89.51	81.94	85.56
Run 3	89.89	81.44	85.46

3 Results and Discussion

We applied the methods discussed above on the CDR data sets provided by the track organisers. For disease name recognition, using a CRF model trained on the CDR Training data set (consisting of 500 PubMed abstracts), we obtained an F-score of 81.70% (precision=87.77% and recall=76.41%) on the CDR Development set (also with 500 abstracts), according to the evaluation library provided.

Three different versions of our disease name normalisation approach have been officially evaluated on the CDR Test corpus of 500 abstracts, the results of which are presented in Table 2. All of them exploited the automatic Greek/Latin medical root/affix translation technique, although using different thresholds in determining whether the translation should be carried out. In the first version (Run 1), the translation is performed only if the string similarity between the mention in question and the topmost candidate is below a threshold of 0.92 (optimised for recall). It made use of a version of MeSH that included only mentions from the CDR Training and Development sets. Both the second (Run 2) and third (Run 3) versions leveraged a MeSH dictionary that additionally incorporated mentions from the NCBI Disease Corpus. A threshold of 0.94 was applied in Run 2 while 0.96 was used in Run 3 (optimised for precision) .

We wrapped our methods as a Web service that accepts and outputs data in the BioC format. It can be accessed at the following URL: `http://nactem.ac.uk/biocreative/dner?format=bioc&run=x`, where x can be any of 1, 2 or 3, depending on the desired version of the normalisation approach, as described above.

References

1. LingPipe 4.1.0. `http://alias-i.com/lingpipe`, accessed: July 2015
2. Medical Prefixes, Suffixes, and Combining Forms. `http://www.stedmansonline.com/webFiles/Dict-Stedmans28/APP05.pdf`, accessed: July 2015
3. NERsuite: A Named Entity Recognition toolkit. `http://nersuite.nlplab.org`, accessed: July 2015
4. Amberger, J., Bocchini, C., Hamosh, A.: A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). Human Mutation 32(5), 564–567 (2011)
5. Banville, D.: Mining chemical structural information from the drug literature. Drug Discovery Today 11(1), 35–42 (2006)

6. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(Database issue), D267–70 (2004)
7. Chang, J.T., Altman, R.B.: Extracting and characterizing gene-drug relationships from the literature. Pharmacogenetics 14(9), 577–586 (Sep 2004)
8. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. pp. 73–78 (2003)
9. Comeau, D.C., Islamaj Doan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BioC: a minimalist approach to interoperability for biomedical text processing. Database 2013 (2013)
10. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J.: The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research (2014)
11. Deftereos, S., Andronis, C., Friedla, E., Persidis, A., Persidis, A.: Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 3(3), 323–334 (2011)
12. Dogan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 47, 1–10 (2014)
13. Jessop, D., Adams, S., Willighagen, E., Hawizy, L., Murray-Rust, P.: OSCAR4: a flexible architecture for chemical text-mining. Journal of Cheminformatics 3(1), 41 (2011)
14. Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., Parkinson, H., Schriml, L.M.: Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research 43(D1), D1071–D1078 (2015)
15. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: CHEMDNER: The drugs and chemical names extraction challenge. Journal of Cheminformatics 7(Suppl 1), S1 (2015)
16. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
17. Leaman, R., Islamaj Doan, R., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 29(22), 2909–2917 (2013)
18. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: Annotating chemicals, diseases, and their interactions in biomedical literature. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Sevilla, Spain (2015)
19. Lipscomb, C.E.: Medical Subject Headings (MeSH). Bull Med Libr Assoc. 88(3), 265–266 (2000)
20. Miwa, M., Saetre, R., Miyao, Y., Tsujii, J.: Protein-protein interaction extraction by leveraging multiple kernels and parsers. International Journal of Medical Informatics 78(12), e39–e46 (2009)
21. Tsuruoka, Y., Tateisi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Advances in Informatics - 10th Panhellenic Conference on Informatics, Lecture Notes in Computer Science, vol. 3746, pp. 382–392. Springer-Verlag, Volos, Greece (November 2005)

22. Wei, C.H., Kao, H.Y., Lu, Z.: PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research 41(W1), W518–W522 (2013)
23. Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C., Lu, Z.: Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Sevilla, Spain (2015)
24. Yeh, A., Morgan, A., Colosimo, M., Hirschman, L.: BioCreAtIvE Task 1A: gene mention finding evaluation. BMC Bioinformatics 6(Suppl 1), S2 (2005)