An Inference Method for Disease Name Normalization

Rezarta Islamaj Dogan and Zhiyong Lu
National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, MD 20894, USA
Rezarta.Islamaj@nih.gov, Zhiyong.Lu@nih.gov

Abstract

PubMed® and other literature databases contain a wealth of information on diseases and their diagnosis/treatment in the form of scientific publications. In order to take advantage of such rich information, several text-mining tools have been developed for automatically detecting mentions of disease names in the PubMed abstracts. The next important step is the normalization of the various disease names to standardized vocabulary entries and medical dictionaries. To this end, we present an automatic approach for mapping disease names in PubMed abstracts to their corresponding concepts in Medical Subject Headings (MeSH®) or Online Mendelian Inheritance in Man (OMIM®). For developing our algorithm, we merged disease concept annotations from two existing corpora. In addition, we hand annotated a separate test set of decease concepts for our method evaluation. Different from others, we reformulate the disease name normalization task as an information retrieval task where input queries are disease names and search results are disease concepts. As such, our inference method builds on existing Lucene search and further improves it by taking into account the string similarity of query terms to the disease concept name and its recognized synonyms. Evaluation results show that our method compares favorably to other state-of-the-art approaches. In conclusion, we find that our approach is a simple and effective way for linking disease names to controlled vocabularies and that the merged disease corpus provides added value for the development of text mining tools for named entity recognition from biomedical text. Data is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Dogan/disease.html

Introduction

Automatic identification of disease names is an important named entity recognition task because disease information is the highest sought non-bibliographic information type in PubMed (Islamaj Dogan et al. 2009) as well as a frequent information type on Google trends (Pelat et al. 2009). Therefore, automatic recognition of diseases mentioned in medical communications and/or biomedical literature is essential not only for improving retrieval of relevant documents, but also for extraction of relevant information pertaining to particular diseases and knowledge discovery. Concept identification and named entity recognition are hot research topics in natural language processing, and much work has been dedicated to identifying biomedical concepts such as disease names, gene/protein names, drugs and chemical names (Jimeno et al. 2008, Leaman et al. 2009, Tanabe et al. 2005, Chowdhury and Lavelli 2010, Smith et al. 2008).

In order to build better models that perform concept identification, considerable effort needs to be dedicated for the development of manually annotated high-quality corpora (Campos et al. 2012, Thomson et al. 2009). In this regard, we recently developed a large and rich disease name corpus containing a set of 793 PubMed abstracts (Islamaj Dogan and Lu 2012). When used as gold-standard data for a state-of-the-art machine learning system, our corpus was able to significantly improve its performance for disease name recognition. The objective of that work was not only to facilitate information retrieval tasks that involve diseases, but also to facilitate future applications of complex information retrieval tasks connecting diseases to treatments, causes or other types of information. Hence, the next important step towards that goal is the entity normalization task that involves mapping mentions to some standard database/ontology identifiers.

Traditionally, several biomedical entity normalization tasks have been explored such as gene/protein normalization (Lu et al. 2011, Wei and Kao, 2011, Huang et al. 2011) and species/organism normalization (Naderi et al. 2011). However, there have been few attempts in the context of disease name normalization (Neveol et al. 2012) perhaps due to the lack of training and evaluation data, and/or the fact that "disease" as a category has a very loose definition and covers a wide range of concepts.

In this study, we present an inference method that aims to map disease names in PubMed abstracts to their corresponding concepts in Medical Subject Headings (MeSH®) and in Online Mendelian Inheritance in Man (OMIM®). More specifically, we propose to address the normalization task in the framework of information retrieval where input

queries are disease names and search results are disease concepts. As such, our inference method builds on existing Lucene search and further improves it by taking into account the string similarity of query terms to the disease concept name and its listed synonyms in the controlled vocabulary.

Our approach differs from existing NLP methods such as Norm[1] and MetaMap (Aronson and Lang 2010), both freely accessible biomedical natural language processing tools, provided by the U.S. National Library of Medicine (NLM). Norm is a tool for addressing the problem of name variation: it produces a normalized version of an input string in lower case, without punctuation or genitive markers, stop words or symbols. The words of the original string are then transformed into their uninflected form and sorted in alphabetical order. In Norm, non-ASCII characters are mapped to ASCII, etc. This particular way of string processing is important for medical concepts, as it swiftly unifies strings such *cancer of the pancreas* and *pancreatic cancer*. MetaMap is a tool which processes biomedical text and returns all the mappings to UMLS concepts, semantic types and more. This intensive, knowledge-based, natural language processing and computational linguistic method lies at the foundation of the Medical Text Indexer, which is used for indexing of biomedical literature at NLM. Note that in this work, UMLS CUIs produced from MetaMap, were translated into their corresponding MeSH ids for our evaluation purposes.

Methods

Data Sources and Preparation

The datasets we used in this study are: the MEDIC disease vocabulary, the development set, and the validation set extracted from the NCBI disease corpus.

The MEDIC Disease Vocabulary

The MEDIC disease vocabulary (Davis et al. 2012) is a recent work of the Comparative Toxicogenomics Database project. MEDIC is a manually curated dataset that for each disease name associates a descriptor from the "Diseases" category of the NLM MeSH resource, or a genetic disorder identifier from the OMIM database. We are using the MEDIC version downloaded on April 17, 2012, which listed a set of 9,661 disease names. In addition it contained synonyms for 91% of the listed diseases and definition strings for 47% of the listed diseases.

The Development Set

The development set is produced from three different sources: the EBI disease corpus (Jimeno et al. 2008), the AZDC disease corpus (Leaman et al. 2009) and the NCBI disease corpus (Islamaj Dogan and Lu 2012). EBI disease corpus contains a set of 856 PubMed sentences extracted from 642 PubMed abstracts manually annotated for UMLS concept identifiers (CUI). This corpus does not identify individual disease mentions in text; instead it has a list of UMLS CUIs corresponding to each sentence. The AZDC disease corpus contains a set of 2,783 PubMed sentences extracted from 793 PubMed abstracts and is manually annotated both for the disease mentions in text and for the corresponding UMLS CUIs. The NCBI disease corpus is a collection of 793 PubMed abstracts that covers both EBI and AZDC disease corpora. The disease name annotations are fully re-examined and re-annotated in this corpus expanding the coverage to include all sentences in all the listed PubMed abstracts.

First, we extracted the set of overlapping sentences in both the EBI and AZDC corpora, for which both corpora agreed on the UMLS CUI annotations. Next, for each UMLS CUI we found its corresponding MeSH identifier. This set of PubMed sentences, annotated with the disease mentions as specified in the NCBI disease corpus, and associated with the list of MeSH identifiers as matched for the UMLS CUIs annotated in the EBI and AZDC corpora, constituted our development set. This collection consists of 516 sentences, from 414 PubMed abstracts, with a total of 1,114 disease mentions that range from 1-7 mentions per sentence, and a total of 841 MeSH identifiers which range from 1-6 identifiers per sentence.

The Validation Set

The validation set is a set of 50 sentences extracted from the NCBI disease corpus so that 1) they belong to different PubMed abstracts than those of the development set; 2) only one sentence is used per each abstract in order to increase diversity; and 3) there is at least one disease mention in each sentence. Two annotators worked on the validation set and assigned a MEDIC identifier to each annotated disease mention. The annotators had access to the PubMed abstracts the sentences were extracted from, the MEDIC disease dictionary as well as MeSH repository and UMLS. The annotation process consisted of two phases: 1) both annotators worked on 10 sentences (annotator agreement 78.6%) after which they discussed their annotations and resolved their differences, and 2) each annotator worked on 20 sentences individually. This set of sentences was used as the validation set for our inference method.

Disease Normalization Methods

We formalize the problem of assigning disease identifiers to medical documents mentioning those disease names with the following abstract representation:

Let D represent a set of document terms, and V represent a set of controlled vocabulary terms. We search for a map-

[1] http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/userDoc/tools/norm.html

ping $D \rightarrow V$ from the set of document terms D to the set of controlled vocabulary terms V so that, if $|D| = N$, and $|V| = M$, then for each D_i, where $i = 1 \ldots N$, we identify a mapping $D_i \rightarrow \{v +\}$, where $v \in V$, and the number of values v associated with D_i is variable for each D_i.

Controlled disease vocabularies usually have a format similar to the following: A vocabulary term v, where v is often referred to as the preferred name, a unique ID, a list of synonyms and a possible definition string describing the vocabulary term. Medical text documents on the other hand, such as PubMed abstracts, contain a number of diseases mentioned in the text, which may also be abbreviated. As such, assuming that we know the list of disease mentions in text, in order to assign appropriate disease concept identifiers, we developed the following integrated method:

First, we built a Lucene search engine (http://lucene.apache.org) to search each disease mention against a disease vocabulary database. Next, we systematically reordered the search results, and inferred the correct disease ID for each disease mention. For abbreviated disease names, we resolved all abbreviated forms of disease mentions into their long form definitions, and used those to infer the concept identifiers. Detailed steps are described below:

Lucene Search

A Lucene search was setup to identify top-1 and top-5 MEDIC identifiers for each unique disease mention in the NCBI disease corpus. The setup specified the search of the disease query using the preferred name, synonyms list, as well as the definition string fields in the MEDIC database. The Lucene score and the corresponding MEDIC identifier were recorded for each search result. MEDIC identifiers can be OMIM identifiers, MeSH identifiers, or both. In order to have a valid comparison, only the MeSH identifiers were considered for the purposes of evaluation when using the development set.

Abbreviation resolution

In PubMed abstracts and all other biomedical literature, abbreviations are a preferred way of referring to disease names. In fact, our previous study on PubMed query logs (Islamaj Dogan et al. 2009) specifically identified that people search for disease names using their abbreviations. However, an abbreviated text string does not correspond to the same unique text string in the biomedical literature. For example: the short form HD may stand for: Huntington's disease, Hansen's disease, hip dysplasia, or Hodgkin's disease. In order to resolve abbreviations to their correct disease concept we followed this process: First, we employed the abbreviation-definition identifier (Yeganova et al. 2011) on all abstracts in the NCBI disease corpus to extract long form–short form *(LF–SF)* pairs. This program has been shown to be highly accurate on biomedical literature

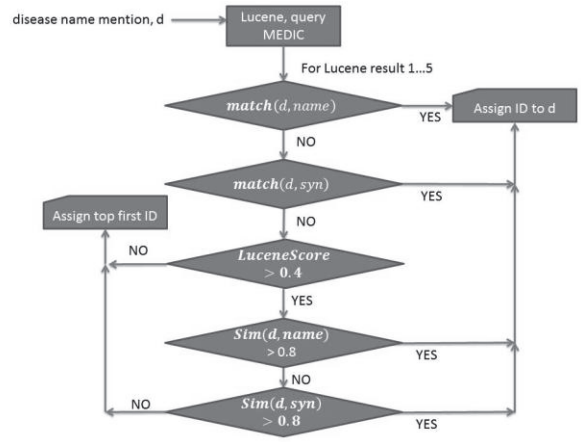

Figure 1 Illustration of the rule-based inference method

text. Second, we resolve each disease mention (in the development set and in the validation set) that contains an abbreviated form to its complete disease name according to the following two rules:

1. Replace trivial abbreviations *(SF)* with their complete long form *(LF)*.

A trivial abbreviation is defined as: a disease mention, *m*, annotated in one of the abstracts of the disease corpus, for which there exists an abbreviation definition, $m \equiv LF$, within the same PubMed abstract, not necessarily within the same sentence, in which *m* is the short form.

2. Resolve non-trivial abbreviations

A non-trivial abbreviation is defined as: m_1, a token in disease mention *m*, annotated in one of the abstracts of the disease corpus, for which there exists an abbreviation definition, $m_1 \equiv LF$, within the same PubMed abstract, not necessarily within the same sentence, in which m_1 is the short form.

Rule-based Inference method

Our rule-based inference method works as described in Figure 1. If a *disease name* query has produced any Lucene results, the following procedural checks need to be performed:

1. Check whether *disease name* matched the MEDIC result preferred name,
2. Check whether *disease name* matched any of the MEDIC result synonym names, if any,
3. Check the Lucene search score
4. Check for any string similarity between *disease name* and the MEDIC result preferred name,
5. Check for any string similarity between *disease name* and any of the MEDIC result synonym names.

String matching

Although string matching and string similarity are fairly common in natural language processing, here we describe how we perform these tasks in our rule-based inference:

1. String matching

We consider two given textual string candidates $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ to match $\boldsymbol{match}(\boldsymbol{s_1}, \boldsymbol{s_2})$ if, after lower-casing, non-alphanumeric character removal and alphabetical word-sorting, both candidates produce the same textual string.

2. String similarity

We consider two given textual string candidates $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ to be similar $\boldsymbol{sim}(\boldsymbol{s_1}, \boldsymbol{s_2})$ if they produce a score of at least 0.8 in a scale from 0 to 1 where 0 means that the strings are entirely different and a 1 means that the strings are identical. The algorithm used is an approximate string matching algorithm which roughly works by looking at the smallest number of edits to change one string into the other (Myers, 1986).

Finally, after the rule-based inference has been applied and the normalization process has been performed for the whole document, each disease mention is re-examined to determine whether it is an abbreviated form. In these cases, the long form definitions of the abbreviated mentions are used to infer the normalization concepts for abbreviated mentions.

Comparison with other methods

In this work we explored several statistical and natural language processing methods for disease name normalization. Here we describe MEDIC dictionary string matching with Specialist lexicon tools, as well as MetaMap processing:

1. Specialist Lexicon string matching

We used Norm to compile a Norm-version of a disease name dictionary that comprised of all entries in the MEDIC dictionary, their synonyms and their corresponding identifiers. The same process was applied to the disease mentions in the development and validation sets. The results of this string matching evaluation are reported as Norm in the results section.

2. MetaMap Processing

In this work, disease mentions were mapped through the MetaMap search engine to corresponding UMLS CUIs. Next, each UMLS CUI was separately mapped to a MeSH identifier, and this set of MeSH identifiers obtained for the development and validation dataset was evaluated against the manual annotations. These results are reported as MetaMap results.

Experimental Design and Evaluation

A standard way of measuring the success of an algorithmic method is to evaluate its performance on a collection of documents which have been labeled beforehand. In our case, we used the development set to refine the inference rules, and the validation set to measure the performance. The inference method is applied to each disease annotated mention in text to produce a list of database concept identifiers from the MEDIC vocabulary, thus creating an automatic mapping. Then the produced mapping was compared to the gold-standard mapping for both development and validation sets.

We evaluated performance using Precision, Recall and F-measure for the produced mappings $D_i \rightarrow \{v +\}$, for each sentence, and then report the average values over all the sentences in the development set, and the validation set. There is a difference in the evaluation results of the validation set compared to those of the developing dataset. The set of gold-standard mappings for the development set was only drawn from MeSH identifiers. As a result, in the event that the inference method produced an OMIM identifier (MEDIC dictionary contains both MeSH and OMIM identifiers), that result was considered a miss during evaluation on the developing set. However, the set of gold-standard mappings for the validation set was produced using MEDIC dictionary as a resource. As such, the whole output of the inference method was considered when evaluating the performance of the inference method on the validation set.

Results

Table 1 shows Lucene search results for the top-1 and top-5 results, which established a recall upper bound (89%) on the performance of our inference method. Table 2 summarizes the inference method results on different development stages. It was during these stages that we refined the rule-based inference method. Our final result marks a recall of 81% with a precision of 76% and an F-measure of 78% on the development set.

Table 3 compares the result of the inference method with two other competing methods, Norm and MetaMap. As shown, our inference method achieves higher recall, and a statistically significant improvement on F-measure.

Finally, when the inference method is applied on the validation set, we achieve an F-measure of 79%, averaging the results on the sentence level. This result is similar to what was obtained for the development set, and suggests that our inference rules are robust on unseen data. It is also interesting to note that even though the development of the inference method was restricted to only MeSH identifiers, the method itself is portable and applicable to other datasets and other dictionary resources. It is a quick and effective way of normalizing disease names in biomedical text.

Conclusions

In this paper, we address the problem of disease name normalization, and we present a solution that links any disease mention annotated in a given PubMed abstract, to a standardized medical vocabulary entry, such as MeSH and OMIM. We used Lucene search engine to identify top-5 results when querying each PubMed disease mention against disease concepts in MEDIC. Then, we use inference to assign the most appropriate MeSH identifiers to the

Table 1 Results of Lucene search on the training dataset. We show the top-5 and top-1 results.

	TOP-5			TOP-1		
Method	Recall	Precision	F-measure	Recall	Precision	F-measure
Default Lucene Search	0.889	0.209	0.340	0.749	0.619	0.679

Table 2 Results of inference method on the training dataset. The size of the developing dataset is 516 sentences. The values of Precision, Recall and F-measure are averaged per sentence on the whole developing dataset.

Inference Method	Recall	Precision	F-measure
$match(d, name)$	0.484	0.550	0.515
$+ match(d, syn)$	0.715	0.780	0.745
$+ Lucene\ score > 4$ and $sim(d, name) > 0.8$	0.750	0.788	0.769
$+ Lucene\ score > 4$ and $sim(d,\ syn) > 0.8$	0.767	0.775	0.771
+ Lucene top 1	0.808	0.759	0.783

Table 3 Results of different methods on the training dataset. We compare METAMAP, MEDIC dictionary string matching with Specialist Lexicon tools, and Inference method.

Method	Recall	Precision	F-measure
Norm	0.714	0.742	0.717
METAMAP	0.738	0.770	0.754
Inference	**0.808**	**0.759**	**0.783**

PubMed disease mention, by effectively re-ranking and eliminating non-relevant Lucene choices. We solve the problem of mapping abbreviated disease mentions to their correct identifiers, by, first, making use of the whole PubMed abstract to identify the abbreviated disease mentions and their long form definition. Next, correct mappings for abbreviated disease mentions are inferred from the high-confidence mappings of their long form definitions. Finally, our evaluation results showed that this method provides MeSH identifiers matching those of manually annotated data, and furthermore, the error analysis revealed that the produced results are not unexpected and unrelated.

In addition, we produced a development and a validation set of selected PubMed sentences with annotated disease mentions and high-quality mappings to MeSH and OMIM identifiers[2], suitable for developing and refining other disease concept normalization methods.

Our future work includes detailed error analysis and creation and release of a manually curated large scale gold-standard dataset that links PubMed disease mentions to MeSH and OMIM identifiers. We are also applying machine learning in matching disease mentions with their standardized medical vocabulary counterpart, in order to improve precision.

[2] Data is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Dogan/disease.html

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Aronson, A., Lang, F. 2010. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3): 229-236.

Campos, D., Matos, S., Lewin, I., Oliveira, J., Rebholz-Schuhmann, D. 2012. Harmonisation of gene/protein annotations: towards a gold standard MEDLINE. *Bioinformatics.*

Chowdhury, F.M., Lavelli, A. 2010. Disease mention recognition with specific features. BioNLP, 91-98.

Davis, A.P., Wiegers, T.C., Rosenstein, M.C., Mattingly, C.J. 2012. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database (Oxford)*. bar065.

Huang, M., Liu, J., and Zhu, X. 2011. GeneTUKit: a software for document-level gene normalization. *Bioinformatics*. 27(7):1032-1033.

Islamaj Dogan, R., Lu, Z. 2012, An improved corpus of disease mentions in PubMed citations, in *BioNLP* 2012.

Islamaj Dogan, R., Murray, G. C., Neveol, A., Lu, Z. 2009. Understanding PubMed user search behavior through log analysis. *Database (Oxford)*: bap018.

Jimeno,A., Jimnez-Ruiz, E., Lee, V., Gaudan, S., Berlanga,R., Reholz-Schuhmann, D. 2008. Assessment of disease named entity

recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3).

Leaman, R., Miller, C., Gonzalez, G. 2009. Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. *Symposium on Languages in Biology and Medicine*, 82-89.

Lu, Z., Kao, H.Y., Wei, C.H., Huang, M., Liu, J., Kuo, C.J., Hsu, C.N., Tsai, R.T., Dai, H.J., Okazaki, N., Cho, H.C., Gerner, M., Solt, I., Agarwal, S., Liu, F., Vishnyakova, D., Ruch, P., Romacker, M., Rinaldi, F., Bhattacharya, S., Srinivasan, P., Liu, H., Torii, M., Matos, S., Campos, D., Verspoor, K., Livingston, K., and Wilbur, W.J., 2011. The Gene Normalization Task in BioCreative III, *BMC Bioinformatics*.

Myers, E., 1986. An O(ND) difference algorithm and its variations, *Algorithmica* 1 (2), pp.251-266.

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

Neveol, A., Li, J., Lu, Z. 2012. Linking Multiple Disease-related resources through UMLS. *ACM International Health Informatics*.

Pelat C., Turbelin C., Bar-Hen A., Flahault A., Valleron A-J. 2009. More diseases tracked by using Google trends. *Emerg Infect Dis*.

Smith L., Tanabe L.K., Ando R.J., Kuo C.J., Chung I.F., Hsu C.N., Lin Y.S., Klinger R., Friedrich C.M., Ganchev K., Torii M., Liu H., Haddow B., Struble C.A., Povinelli R.J., Vlachos A., Baumgartner W.A. Jr., Hunter L., Carpenter B., Tsai R.T., Dai H.J., Liu F., Chen Y., Sun C., Katrenko S., Adriaans P., Blaschke C., Torres R., Neves M., Nakov P., Divoli A., Maña-López M., Mata J., Wilbur W.J. 2008.Overview of BioCreative II gene mention recognition. *Genome Biology*, 9 Suppl 2:S2.

Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.J. 2005. GENETAG: a tagged corpus for gene /protein named entity recognition. *BMC Bioinformatics*, 6:S3.

Thompson, P., Iqbal, S.A., McNaught, J., Ananiadou, S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349.

Wei, C-H. and Kao, H-Y. 2011. Cross-species gene normalization by species inference. *BMC Bioinformatics*. 12(S8):S5.

Yeganova, L., Comeau, D.C., Wilbur, W.J. 2011. Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC Bioinformatics*. S3:S6.

Yeh, A., Morgan, A., Colosime, M., Hirschman, L. 2005. BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2.