LeadMine: Disease identification and concept mapping using Wikipedia

Daniel M. Lowe*[1], Noel M. O'Boyle[2], Roger A. Sayle[3]

NextMove Software Ltd, Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge, United Kingdom

*[1]daniel@nextmovesoftware.com;
[2]noel@nextmovesoftware.com;
[3]roger@nextmovesoftware.com

Abstract. LeadMine, a dictionary/grammar based entity recognizer, was used to recognize and normalize both chemicals and diseases to MeSH IDs. The lexicon was obtained from 3 sources: MeSH, the Disease Ontology and Wikipedia. The Wikipedia dictionary was derived from pages with a disease/symptom box, or those where the page title appeared in the lexicon. Composite entities (e.g. heart and lung disease) were detected and mapped to their composite MeSH IDs. For chemical-induced disease relationships we developed a simple pattern-based system to find relationships within the same sentence. If none of our patterns matched the abstract, a heuristic was applied in which the most likely chemical/s were associated with all diseases. Diseases unlikely to be caused by a chemical were removed. The MeSH hierarchy was used to remove redundant relationships. Our system achieved F1-scores of 86.12%, for disease concept ID recognition, and 52.20% for chemical-induced disease relationships, on the test set.

Keywords. LeadMine; CDR; Wikipedia

1 Introduction

Identifying the relationships between chemicals and diseases has many applications in biomedical research and healthcare. The CDR (Chemical-Disease Relation) challenge was organized to encourage research into this area and evaluate current solutions. The challenge was formed of two subtasks; the first was to identify diseases and normalize them to MeSH (Medical Subject Headings) IDs. The second was to identify causal relationships between chemicals and diseases,

with the results reported as MeSH ID pairs. Further information about the challenge is available in the challenge task papers[1, 2].

2 Discussion

To facilitate mapping of entities to MeSH IDs we used a dictionary based approach. The dictionary was derived from three sources: MeSH[3], the Disease Ontology[4] and Wikipedia[5].

2.1 MeSH

The "MeSH descriptors and qualifiers" and the "Supplementary Concept Records" files were downloaded. From these all terms (and synonyms thereof) which were in MeSH trees C (Disease) or F03 (Mental Disorders) were extracted with their MeSH ID mapping. By special case the following tree branches were excluded C23.550.291 (Disease attributes), C23.550.260 (Death) and C26 (Wounds and Injuries (unspecified)). These branches were excluded either because the concept was too vague or because the concept isn't by some definitions a disease. MeSH supplementary records were selected if they referred to a disease MeSH ID (as determined by the aforementioned criteria).

2.2 Disease Ontology

The Disease Ontology was downloaded in OBO format and concept titles and their synonyms were extracted. Where a cross-reference to MeSH was present these terms were associated with the corresponding MeSH ID.

2.3 Wikipedia

A dump of current Wikipedia page articles (enwiki-20150602-pages-articles.xml.bz2) was downloaded. Pages with disease or symptom boxes that contained a link-out to MeSH were identified. From these the page title and all redirects to the page were recorded as mapping to that MeSH ID. Occasionally a MeSH tree number was used instead of an ID requiring these to be converted to the corresponding ID. A large collection of terms to ignore and a small collection of page titles to ignore were empirically assembled. Examples of terms to ignore include

"Allergy medication", "HPV test", "History of acne", "Rehydrated", "2009 flu pandemic". These highlight the problem that while Wikipedia pages are a very rich source of synonyms (especially common names and adjectival forms) that the redirects are not semantic. Pages ignored did not relate to a disease e.g. MUMPS (a programming language).

Additionally if a page title in Wikipedia matched a term in the dictionary assembled from MeSH and the Disease Ontology, it was assumed to relate to the same concept and hence all redirects to the page were linked to the MeSH ID used by the term in the dictionary.

2.4 Final dictionary preparation

A final dictionary was assembled by adding the source dictionaries in the following order: manually curated dictionary (mostly used to correct MeSH IDs referenced from Wikipedia terms), MeSH terms, Disease Ontology terms, terms taken from the training/development corpus and terms taken from Wikipedia. Spelling variants e.g. tumor vs tumour, were generated at the point of adding a term to the dictionary. If a term appeared in two source dictionaries the ordering determined the MeSH ID used in the final dictionary. A stop word list including mostly short abbreviations, gene names and disease names with their abbreviation, was used to remove unwanted terms.
Index names e.g. "Abnormality, Congenital" were uninverted by splitting on comma space and rearranging. Cases where a list was intended (e.g. ", or") were left unchanged. Qualifiers (e.g. ", with") were moved to the end of the term with the comma removed. Finally terms that may be synonyms were generated e.g. "infection" replaced by "disease", "cancer" replaced by "carcinoma".

2.5 Disease recognition with LeadMine

LeadMine[6] was configured to use this dictionary for recognition and normalizing recognized entities to MeSH IDs. A low level of spelling correction was used to recognize minor spelling errors. After recognition composite entities were detected e.g. "heart and lung disease", and mapped to MeSH IDs corresponding to the reconstructed entities i.e. "heart disease" and "lung disease". By special case where MeSH dis-

tinguishes the drug-induced form of a disease, the MeSH ID for the drug-induced form was always chosen.

2.6 Effect of adding Wikipedia dictionary

By adjusting the source dictionaries, the performance change of including the Wikipedia terms was quantified on the development set. Our final system corrected some of the mistakes in the Wikipedia terms e.g. heart disease linked to the MeSH ID for cardiovascular disease.

Dictionaries	Precision	Recall	F_1-score
Wikipedia	79.3	61.3	69.1
MeSH/Disease Ontology	91.6	67.1	77.4
MeSH/Disease Ontology/Wikipedia	85.1	73.1	78.6

2.7 Chemical-induced disease relationship extraction

LeadMine was used to detect chemical entities using a configuration similar to that used in the CHEMDNER-Patents task with the exception of an additional dictionary of special cases noted in the annotation guidelines e.g. oral contraceptive. The terms in the chemical branch of MeSH were used to resolve recognized terms to MeSH IDs. Where an exact match was not found variants were tried e.g. plural of recognized term. This achieved an F_1-score of 92.3% for chemical MeSH ID recognition on the development set.

Sentence detection and part of speech tagging were performed by OpenNLP[7]. The part of speech tags were used to groups diseases/chemicals into blocks by grouping all entities not separated by a verb or, preposition or subordinating conjunction. Patterns were used to identify relationships between chemical and disease groups. Most patterns were regex-based typically consisting of attempting to find a key word/phrase e.g. chemical <caused> disease, where caused is:

```
.*(-associated|(?<!not |[a-z])(associated with|cause[sd]|...))(?! no ).*
```

As can be seen from the pattern, a simple attempt is made to avoid identifying negative associations. The following table summarizes the patterns used and their performance when evaluated on the union of the training and development sets.

Patterns where the chemical preceded the disease:

Pattern	True Positives	False Positives	Precision
Chemical <caused>	528	219	70.7%
Chemical Disease	41	25	62.1%
Chemical <related to>	8	2	80.0%
<negative effects caused by> chemical	4	2	66.7%
<relationship between> chemical <and>	2	1	66.7%

Patterns where the chemical followed the disease:

Pattern	True Positives	False Positives	Precision
Disease <caused by>	208	79	72.47%
Disease <after or during>	108	76	58.70%
Disease <after or while taking>	73	36	67.00%
Disease <in person taking>	18	4	81.80%
Disease <effect of>	14	14	50.00%
Disease <related to>	14	6	70.00%
Disease <complications of>	12	5	70.60%
<induction of> Disease <by or with>	2	1	66.70%

Patterns were developed by taking a sentence containing a chemical and disease known to be in a chemical-induced disease relationship, and manually identifying the key word/phrase that indicated the relationship. This gives the prototypical relationship pattern which is then expanded by identifying and postulating other synonymous phrases. The "actual" precision of patterns is likely to be underestimated due to the requirement that only the most specific relationship be annotated. This means that if the more specific relationship is not found the less specific relationship is counted as a false positive.

All diseases/chemicals in a group linked by one of these patterns were identified as being in chemical-induced disease relationships (CIDs). When no patterns matched an abstract, optionally, a heuristic is applied to find likely relationships. All chemicals in the title (or failing that the

first most commonly mentioned chemical in the abstract) are associated with all diseases in the entire abstract.

Optionally a filtering step was performed. A small number of diseases were blocked: D064420, D010300, D003643, D066126 and D020258, typically as they were too vague. Additionally a disease's corresponding MeSH tree numbers were used to block C02 (Virus Diseases) and C16.320 (Genetic Diseases, Inborn) as these are unlikely to be caused by chemicals.

In all cases MeSH tree numbers were used to identify redundant relationships i.e. those in which the tree numbers of a disease are entirely refinements of those used in another relationship.

3 Evaluation

We setup Tomcat on an Amazon Web Services instance. The instance had 2 GiB of RAM and 1 core. Disease named entity recognition (DNER) was evaluated on the agreement between the system's MeSH IDs and those in the test set. For chemical-induced disease relationships (CID) our 3 runs correspond to the pattern based system, that system plus filters to improve precision, and the aforementioned system plus a heuristic to find the most likely chemical-disease relationship.

Task	Precision	Recall	F_1-score	Response time
DNER	86.08%	86.17%	86.12%	45.0 ms
CID (pattern-based)	57.65%	36.77%	44.90%	96.9 ms
CID (pattern-based with filters)	60.99%	35.93%	45.22%	121.8 ms
CID (pattern-based with filters and recall increasing heuristic)	52.62%	51.78%	52.20%	119.3 ms

Due to LeadMine's speed of annotation the response time for DNER is likely to be primarily limited by internet latency. To simplify implementation the DNER configuration performed both chemical and disease recognition.

REFERENCES

1. Wei CH, Peng Y, Leaman R, et al. (2015) Overview of the BioCreative V Chemical Disease Relation (CDR) Task. Proceedings of the fifth BioCreative challenge evaluation workshop

2. Li J, Sun Y, Johnson R, et al. (2015) Annotating chemicals, diseases, and their interactions in biomedical literature. Proceedings of the fifth BioCreative challenge evaluation workshop

3. Lipscomb CE (2000) Medical subject headings (MeSH). Bulletin of the Medical Library Association 88:265.

4. Schriml LM, Arze C, Nadendla S, et al. (2012) Disease Ontology: a backbone for disease semantic integration. Nucleic acids research 40:D940–D946.

5. (2015) Wikipedia, the free encyclopedia. https://en.wikipedia.org/. Accessed 31 Aug 2015

6. Lowe DM, Sayle RA (2015) LeadMine: A grammar and dictionary driven approach to entity recognition. Journal of Cheminformatics 7:S5.

7. Apache OpenNLP. http://opennlp.apache.org/. Accessed 31 Aug 2015