BioCreative V CDR Task - Data Annotation Guidelines

Draft, March 16, 2015

【Outline】

Task Description

A text-mining challenge task of automatic extraction of chemical-disease relations (CDR) from PubMed articles was recently proposed in conjunction with BioCreative V (http://www.biocreative.org/). Motivated by real-world needs in drug discovery, biocuration and pharmacovigilance, the CDR task is aimed to both advance text-mining research on relationship extraction and provide practical benefits to various domain applications (e.g. assisting biocuration).

There are two specific subtasks in CDR: (A) Disease Named Entity Recognition and Normalization (DNER). (B) Chemical-induced diseases relation extraction (CID). In both cases, participating teams will be provided with annotated PubMed articles for system training and development.

This document contains the guidelines we will use to prepare the task data for both tasks including (i) marking up disease mentions and assigning associated concept ids (ii) marking up chemical mentions and assigning associated concept ids and (iii) annotating pairs of disease-chemical relations via respective concepts ids. To maximize data interoperability among the BioNLP community, our guidelines are crafted to be as consistent as possible with previous works.

Task Data

Disease/Disorder terminology: The 'Diseases' [C] branch of MeSH 2015, including signs and symptoms.
Chemical terminology: The 'Drugs and Chemicals' [D] branch of MeSH 2015.
Articles: PubMed abstracts are collected from the curated articles in the Comparative Toxicogenomics Database (CTD [1]).

Annotation Tool

Human annotation of disease and chemicals is performed using PubTator [2] (See Figure 1). To accelerate human annotation [3], text-mined disease and chemical results are pre-computed and displayed to the human annotators using DNorm [4] and tmChem [5]. When necessary, the human annotators add new annotations, delete or edit existing ones based on their judgment. They are encouraged to use public resources such as UMLS or Wikipedia to facilitate the annotation process.

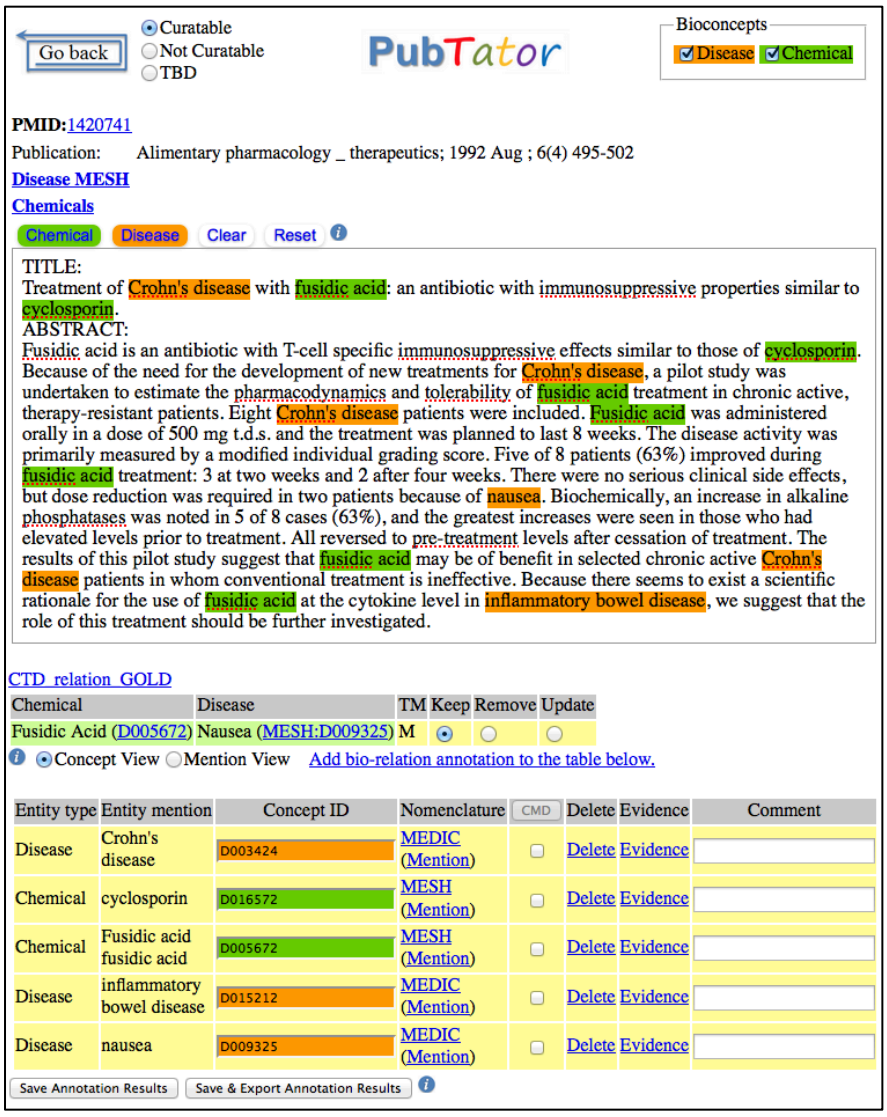

Figure 1. A screenshot of PubTator annotation page.

Guidelines for disease mention and concept id annotation

Please note that whenever possible, we will follow closely the guidelines of constructing NCBI disease corpus for annotating disease mentions [6] and its associated concepts [7].

What to annotate?

1. Annotate the most specific disease mentions and select the best-matching MeSH ID. For instance, the complete phrase "partial seizures" should be preferred over "seizures" as it is more specific.

2. Annotate minimum necessary text spans for a disease. For example, select "hypertension" instead of "sustained hypertension."

3. Annotate all mentions of a disease entity in an abstract. All occurrences of the same disease mention should be marked, including duplicates within the same sentence. PubTator provides an automatic function to help mark the duplicate mentions via a single click.

4. Annotate mentions with morphological variations such as adjectives. For instance, "hypertensive" is annotated as "hypertension."

5. Annotate abbreviations. Abbreviations should be annotated separately. For instance, "Huntington disease (HD)" should be separated into two annotations: "Huntington disease" and "HD" with the same concept id.

6. Annotate all concepts in a composite disease mention using the ''|'' separator. Composite mentions such as: ''ovarian and peritoneal cancer'' should be normalized to the collection of the individual constituents as D010051 (Ovarian Neoplasms) | D010534 (Peritoneal Neoplasms). When possible, we also spell out the individual constituents with their associated text. That is, for the above example, we also provide "ovarian … cancer" for D010051 and "peritoneal cancer" for D010534.

7. Annotate a disease mention using multiple concepts to logically describe the disease mention as a whole, using the ''+'' concatenator. It may be required to represent a disease mention with multiple concepts which are not inherited with each other. For example, "bone marrow oedema" implies Bone Marrow Diseases (MESH: D001855) and Edema (MESH:D004487). Thus, the complete phrase should be annotated as D001855+D004487. When possible, we also spell out the individual constituents with their associated text. That is, for the above example, we also provide "bone marrow" for D001855 and "edema" for D004487.

8. When a disease cannot be normalized, "-1" is used (e.g. erythroblastocytopenia).

What not to annotate?

1. Do NOT include species names as part of a disease. Organism names such as "human" are generally excluded from the preferred mention unless they are critical part of a disease name. Viruses, bacteria, and other organism names are not annotated unless it is clear from the context that the disease is caused by these organisms. e.g. "HIV-1-infected" means the disease caused by the organism "HIV". Thus, "HIV" should be included.

2. DO NOT annotate overlapping mentions. For example, the phrase "hepatitis B virus (HBV) infected" was annotated as one single disease D006509 (Hepatitis B).

3. DO NOT annotate general terms such as: disease, syndrome, deficiency, complications, etc. However, terms such as pain, cancer, tumor and death should be retained.

4. DO NOT annotate references to biological processes such as "tumorigenesis" or "cancerogenesis".

Special cases of disease annotations

1. Rules for "toxicity" related mentions. In this dataset, 'toxicity' and related terms are common. Below are our rules for annotating them.
   a) When the general term 'toxicity' appears by itself, use concept id D064420 (Drug-Related Side Effects and Adverse Reactions).
   b) When a specific type of 'toxicity' is mentioned (e.g. cardiotoxicity or liver toxicity), use concepts under D064420 when possible (e.g. D066126 for 'cardiac toxicity' and D056486 for 'liver toxicity').
   c) When no matching MeSH terms can be found under D064420 for a specific type of toxicity, use the corresponding disease ids. For instance, use D014786 (vision disorders) for 'visual toxicity' and D006311 (hearing disorders) for 'auditory toxicity' and 'ototoxicity.'

2. If the article indicates that a disease is drug induced (again common in this dataset), then one should always try to annotate the disease using terms under D064420 (Drug-Related Side Effects and Adverse Reactions) whenever possible. For instance, 'dyskinesia' is a movement disorder and has its own MeSH ID (D020820). However, when it is clear from the article that 'dyskinesia' is caused by a drug (e.g. PMID: 6727060), then we should use D004409 regardless whether 'drug-induced' is included in the disease mention or not. Other common examples include D056486 (Drug-induced Liver Injury) for 'hepatitis'; D017109 (Akathisia, Drug-induced) for 'akathisis.'

Guidelines for chemical mention and concept id annotation

Please note that whenever possible, we follow closely the guidelines of constructing CHEMDNER corpus for annotating chemical mentions [8]. The basic rule for chemical entity annotation is that the chemical should have a specific structure.

What to annotate?

1. Annotate the most specific chemical mentions and select the best-matching MeSH concept ID. Chemicals which should be annotated are listed as follows:
   a) Chemical Nouns convertible to:
      -A single chemical structure diagram: single atoms, ions, isotopes, pure elements and molecules such as: Calcium (Ca), Iron (Fe), Lithium (Li), Potassium (K), Oxygen (O_2),
      -A general Markush diagram with R groups such as: Amino acids

   b) General class names where the definition of the class includes information on some structural or elemental composition such as: steroids, sugars, fatty acids, saturated fatty acids …

   c) Small Biochemicals
      - Monosaccharides, disaccharides and trisaccharides: Glucose, Sucrose …
      - Peptides and proteins with less than 15 aminoacids: Angiotensin II …
      - Monomers, dimmers, trimmers of nucleotides: e.g. ATP, cAMP …
      - Fatty acids and their derivatives excluding polymeric structures. e.g. Cholesterol, glycerol, prostaglandin E1 …

   d) Synthetic Polymers such as: Polyethylene glycol

   e) Special chemicals having well-defined chemical compositions.
      E.g. “ethanolic extract of Daucus carota seeds (DCE)” in 16755009; “grape seed proanthocyanidin extract” in 11334364.

2. Annotate all mentions of a chemical entity in an abstract.

3. Annotate abbreviations. Some abbreviations are ambiguous by convention. Take “Nitric Oxide (NO)” as an example, “NO” could also be interpreted as a negative response. Ambiguity should be avoided using context.

4. Annotate chemicals when they cannot be normalized. In such a case, “-1” will be used as the concept id.

What not to annotate?

1. DO NOT annotate other terms different from chemical nouns. Adjective forms of chemical names are also excluded. For instance, muscarinic, adrenergic and purinergic in 17244258.

2. DO NOT annotate chemical nouns named for a role or similar, that is, nonstructural concepts (e.g. anti-HIV agents, anticonvulsants, anticholinesterase drug, antipsychotic, anticoagulant, etc).

3. DO NOT annotate very nonspecific structural concepts. e.g. Atom, Ion, Molecular, Lipid, Protein …

4. DO NOT annotate words that are not chemicals in context, even if they are co-incidentally the same set of characters (synonyms and metaphors). For instance, "Gold" should not be annotated if it appears in "gold standard."

5. DO NOT annotate Biomolecules/Macromolecular biochemicals: not large oligomeric and polymeric or established DNA/RNA/protein sequences. Proteins, polypeptides (> 15aa), nucleic acid polymers, polysaccharides, oligosaccharides and other biochemical are excluded. E.g. Insulin, DNA, mRNA, collagen, starch, cellulose, glycogen, lipopolysaccharide, glucocorticoid, glucagon (29 peptides), prolactin (199 peptides) …

6. DO NOT annotate general vague compositions. For instance, according to Wikipedia, the term opiate describes any of the narcotic opioid alkaloids found as natural products in the opium poppy plant, Papaver somniferum, and thus should be excluded.

7. DO NOT annotate special words not to be labeled by convention (e.g. Water, saline, juice, etc).

Special cases of chemical annotations

- Antidepressive Agents (D000928), Estrogens (D004967), DContraceptives, Oral (D003276) are annotated in order to match CTD's relation annotations.

- For combo drugs, mark as one entity and use their corresponding id in MeSH when possible (as opposed to annotating its component separately). For example, select levodopa/carbidopa in PMID:2265898 and assign MeSH ID:C009265 (carbidopa, levodopa drug combination).

Guidelines for annotating chemical-induced disease relations

The chemical-induced disease relation pairs are annotated as part of the Comparative Toxicogenomics Database (CTD) curation (see [9, 10] for details). For the CDR task, some additional updates are performed such that:

1. The annotated relationship includes primarily mechanistic relationships between a chemical and disease. Occasional biomarker relations are also included (e.g. relation between D006719 (Homovanillic Acid) and D006816 (Huntington Disease) in PMID:6453208).

2. The relation should be explicitly mentioned in the abstract.

3. Use the most specific disease in a relationship pair.

References

1. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ: **Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks**. *Nucleic Acids Res* 2009, **37**(Database issue):D786-792.
2. Wei C-H, Kao H-Y, Lu Z: **PubTator: a Web-based text mining tool for assisting Biocuration**. *Nucleic Acids Res* 2013, **41**(W1):W518-W522.
3. Neveol A, Islamaj Dogan R, Lu Z: **Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction**. *J Biomed Inform* 2011, **44**(2):310-318.
4. Leaman R, Islamaj Dogan R, Lu Z: **DNorm: disease name normalization with pairwise learning to rank**. *Bioinformatics* 2013, **29**(22):2909-2917.
5. Leaman R, Wei C-H, Lu Z: **tmChem: a high performance approach for chemical named entity recognition and normalization**. *Journal of Cheminformatics* 2015, **7**(Suppl 1):S3.
6. Dogan R, Lu Z: **An improved corpus of disease mentions in PubMed citations**. In: *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012): 2012; Montreal, Canada*. 91-99.
7. Dogan RI, Leaman R, Lu Z: **NCBI disease corpus: A resource for disease name recognition and concept normalization**. *J Biomed Inform* 2014, **47**:1-10.
8. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe D *et al*: **The CHEMDNER corpus of chemicals and drugs and its annotation principles**. *Journal of Cheminformatics* 2015, **7**(Suppl 1):S2.
9. Davis AP, Wiegers TC, Roberts PM, King BL, Lay JM, Lennon-Hopkins K, Sciaky D, Johnson R, Keating H, Greene N *et al*: **A CTD-Pfizer collaboration: manual**

curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database (Oxford)* 2013, **2013**:bat080.

10. Davis AP, Wiegers TC, Rosenstein MC, Murphy CG, Mattingly CJ: **The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database**. *Database (Oxford)* 2011, **2011**:bar034.