Mining ClinicalTrials.gov via CTTI AACT for drug interventions

Introduction

Drug named entity recognition (D-NER) is a critical step for complex biomedical NLP tasks (8). D-NER is the task of locating drug entity mentioned in unstructured medical texts and classifying their predefined categories (9). NER in biomedicine is focused proteins, genes, diseases, drugs, organs, DNA sequences, RNA sequences and others (10). By mining ClinicalTrials.gov using the Clinical Trial Transformation Initiative – Aggregate Analysis of ClinicalTrials.gov database (ctti-aact db) from Duke University, new and distinctive drug target knowledge can be inferred from drugs and diseases. Extraction of chemical named entities (CNEs)is an essential task since it allows using the obtained data for building chemical-target associations and other relationships relevant for drug discovery.

Method

The AACT contains information about clinical studies registered in ClinicalTrials.gov. From aact db, a static pipe delimited copy of the db was obtained on 31 Oct 2022. Although the downloaded data comprises of 51 tables, only a handful of these are used in the analysis. They include the following five (5) study-related named tables in the AACT database. Once drugs and attributes are extracted, the subsequent step is to link drug names to the corresponding attributes. The tables are also briefly described. Here, I applied a simplistic approach to NER by directly matching textual expressions found in a relevant lexical repository (PubChem) against raw text (drug names) obtained from aact.

- Interventions.txt - exposures (drugs, medical devices, procedures, vaccines, and other products)
- Studies - Basic info about study, including study title, date study registered with ClinicalTrials.gov, date results first posted to ClinicalTrials.gov, dates for study start and completion, phase of study, enrollment status, planned or actual enrollment, number of study arms/groups, etc.
- study_references - Citations to publications related to the study protocol and/or results. Includes PubMed Unique Identifier (PMID) and/or full bibliographic citation.
- conditions- Name(s) of the disease(s) or condition(s) studied in the clinical study, or the focus of the clinical study. Can include NLM's Medical Subject Heading (MeSH)-controlled vocabulary terms.
- browse_condition - list of standard MeSH terms that describe the intervention(s) being addressed by the clinical trial

Interventions.txt

A table with 735,368 rows × 5 columns

Attributes of this table - (id (#), nct_id(e.g., NCT03309709), intervention_type (Behavioral, Biological, Combination Product, Device, Diagnostic Test, Dietary Supplement, Drug, Genetic, Other, Procedure or Radiation), name (e.g., Tramadol), description (e.g., IV 4 mg/ mL tramadol solution into 100 mL normal saline)

Using pandas' df['intervention_type'].value_counts(), the count of intervention_type=="Drug" is 327,906 out of the 735,368). But how many unique ids are there?

Using Linux commands

- Obtain unique ids (uniq nct_id_drugs > unique_drug_ids)
- Obtain the number of unique drug ids (uniq nct_id_drugs|wc -l > 185,768)

This implies that multiple drug names are connected to a study.

Studies.txt

This table was queried using the drug interventions nct_id (uniq nct_id_drugs) using the command - grep -Fwf unique_drug_ids studies.txt > unique_drug_ids_studies. The output is a selection from studies.txt that are only linked to drug interventions. The extracted file consists of 167,771 (# unique studies identified by nct_id) × 69 columns. There are 69 columns, but we are interested in the columns that would enable us to know the distributions of studies by start year (the year the study started), what phase these drug trials are in and their overall status (current state or state of the trial).

Study_references

There are 234,930 references associated with these 167,771 studies. We extracted, from study_references only references that are linked to the unique drug nct_ids. Command used -

```
grep -Fwf unique_drug_ids study_references.txt > unique_drug_ids_study_references
```

Conditions using browse_conditions.txt

We identified conditions that are connected to drug interventions. There are 3674 MeSH terms describing these conditions.

Using command - grep -Fwf unique_drug_ids browse_conditions.txt > unique_drug_ids_conditions_browse_conditions

Identifying drug names

Here are extracting names of drugs from interventions.txt using the unique_drug_ids list generated.

```
grep -Fwf unique_drug_ids interventions.txt > unique_drug_interventions
```

Extract the “name” column - drug name column, convert the characters to lower case; then save to a file.

```
df2 = df['name'].str.lower()
```

```
df2.to_csv('drug_name', index = False)
```

With $ tr -cs "[:alpha:]" "\n" < drug_name > drug_names_tokens, a list of the words was created, one per line, where a word is taken to be a maximal string of letters.

With $ grep -oE '[[:alpha:]]+' drug_names_tokens | grep -vwFf words.txt | sort -f | uniq -ic | sort -nr > drug_names_tokens2, refined list of drug_names_tokens were created with their frequency of occurrence (mentions).

Detailed explanation is here provided - how it works

To get a word count: $ grep -oE '[[:alpha:]]+' text.txt | sort | uniq -c | sort -nr

- grep -oE '[[:alpha:]]+' text.txt (returns all words, minus any spaces or punctuation, with one word per line)
- sort (sorts the words into alphabetical order)
- uniq -c (counts the number of times each word occurs. For uniq to work, its input must be sorted.)
- sort -nr (sorts the output numerically so that the most frequent word is at the top.)

But I want the command to be able to handle mixed cases

```
$ grep -oE '[[:alpha:]]+' input.txt | sort -f | uniq -ic | sort -nr > output.txt
```

- added the -f option to sort so that it would ignore case and
- Added the -i option to uniq so that it also would ignore case

I also want to exclude English words

- so, I pass an English dictionary, words.txt (downloaded from https://raw.githubusercontent.com/dwyl/english-words/master/words_alpha.txt)

$ grep -oE '[[:alpha:]]+' drug_names_tokens | grep -vwFf words.txt | sort -f | uniq -ic | sort -nr > drug_names_tokens2 (10)

SMILES

Query Chemical Identifier Resolver

The R package webchem "Chemical Information from the Web" interacts with a suite of web services for chemical information such as Chemical Identifier Resolver, ChEBI, ChemIDplus, ChemSpider, PubChem, Wikidata etc. Current version is Version 1.2.0. (12)

```
# Multiple inputs (input for this work was 21,667 entities)
```

Example:

```
comp <- c("Triclosan", "Aspirin")

cir_query(comp, "cas", match = "first")
```

Results & Conclusions

Chemical NER

Intervention names of drug trials: 185,768; drug names: 21,667; SMILES: 5,329 with many non-structure terms. The top 5 drug mentions are - 3348 paclitaxel, 3070 cyclophosphamide, 2985 cisplatin, 2620 carboplatin, 2384 gemcitabine. Other drug names mentioned in thousands are - 2224 metformin, 2074 docetaxel, 2041 bevacizumab, 1763 rituximab, 1592 doxorubicin, 1579 capecitabine, 1556 pembrolizumab, 1538 bupivacaine, 1408 oxaliplatin, 1296 fludarabine, 1281 etoposide, 1221 dexmedetomidine, 1209 azd, 1194 nivolumab, 1160 propofol, 1121 irinotecan, 1064 ketamine, and 1042 cytarabine.

MeSH terms

The top MeSH terms mentioned connected to these drugs are neoplasms 47766, neoplasms by site 28070, pathologic processes 26197, neoplasms by histologic type 24750, immune system diseases 20105, digestive system diseases 16956, nervous system diseases16667, cardiovascular diseases 16586, infections 16417, respiratory tract diseases16038, lung diseases 13275, vascular diseases 3086, mental disorders 11992, metabolic diseases11603, endocrine system disease 11522, skin diseases 11333 and neoplasms, glandular and epithelial 11296.

SMILES

Resolving the 21, 667 drug names was done using webChem. 5,329 SMILES were generated from these names.

Drug Target identification

One of the failures of this approach due to limitations posed by the tools deployed. It was not possible to find a way around mapping the chemical identification id(cid) generated from PubChem via webChem to ChEMBL.

Challenges

Drugs are known by any of these three names - chemical name, generic name, or brand name. Chemical name is usually complex; a brand name may not identify a drug (especially once the relevant patents expire) and; generic name is negotiated when the drug is approved for use, but this is regulated by the World Health Organization (WHO) and local agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency

Identifying chemical names is especially challenging for the reasons stated below:

- variety of language expressions
- variants can include an innumerable mix of typographical variants, alternating uses of hyphens, brackets, spacing, and word order
- have many synonyms e.g. triphenyl phosphate, TPP, TPHP, OP(OC6H5)3, phosphoric acid, triphenyl ester, CAS: 115-86-6, ((PhO)3PO)) refer to the same chemical
- mentions can refer to mixtures of chemicals
- often ambiguous, especially when abbreviated
- the complexity makes it difficult to identify exact boundaries for entity mentions and word tokens

ChEMBL or ChEMBL db is a manually curated chemical database of bioactive molecules with drug-like properties. It is maintained by the European Bioinformatics Institute (EBI), of the European Molecular Biology Laboratory (EMBL), based at the Wellcome Trust Genome Campus, Hinxton, UK (11).

Limitations of ChEMBL

- direct string-matching overlooks term ambiguity and variability (2)
- different tokens may refer to the same semantic type (term variability) (1).
- entities differ in length and may contain more than one token, each token was annotated to capture information
- ChEMBL records only link from PubChem via PubChem Bioassay (AIDs) but lacks direct links to PubChem substances (SIDs) (4).

Deploying Statistical learning models for NER is one of the common approaches in overcoming some of these challenges

This approach defines drug NER as a classification task that can classify an input expression either as an entity or not; and requires annotated data for training supervised models with human curated data, the gold-standard.

But biomedical gold or silver standard corpus (3) is required to make this approach a success.

References

1. A.M. Cohen, W.R. Hersh A survey of current work in biomedical text mining Brief Bioinform, 6 (1) (2005), pp. 57-71

2. S. Ananiadou, J. McNaught Text Mining for Biology and Biomedicine Artech House, Inc., Norwood, MA, USA (2005)

3. Assessment of NER solutions against the first and second CALBC silver standard corpus

4. N. Collier, U. Hahn, D. Rebholz-Schuhmann, F. Rinaldi, S. Pyysalo (Eds.), Semantic Mining in Biomedicine vol. 714 of CEUR Workshop Proceedings, CEUR-WS.org (2010)

5. Muresan S, Sitzmann M, Southan C. Mapping between databases of compounds and protein targets. Methods Mol Biol. 2012; 910:145-64. doi: 10.1007/978-1-61779-965-5_8. PMID: 22821596; PMCID: PMC7449375.

6. Semeval-2013 task 9: extraction of drug–drug interactions from biomedical texts (DDIExtraction 2013)

7. S. Manandhar, D. Yuret (Eds.), Second Joint Conference on Lexical and Computational Semantics (*SEM), vol. 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia, USA (2013 June), pp. 341-350

8. Korkontzelos I, Piliouras D, Dowsey AW, Ananiadou S. Boosting drug named entity recognition using an aggregate classifier. Artif Intell Med. 2015 Oct;65(2):145-53. doi: 10.1016/j.artmed.2015.05.007. Epub 2015 Jun 17. PMID: 26116947.

9. Zeng, D.; Sun, C.; Lin, L.; Liu, B. LSTM-CRF for Drug-Named Entity Recognition. Entropy 2017, 19, 283. https://doi.org/10.3390/e19060283

10. https://unix.stackexchange.com/questions/316556/find-n-most-frequent-words-in-a-file-with-a-stop-words-list-from-the-command-lin

11. https://pubchem.ncbi.nlm.nih.gov/source/ChEMBL

12. Szöcs E, Stirling T, Scott ER, et al (2020) webchem: An R Package to Retrieve Chemical Information from the Web.