Use of ontologies, such as UNII, to link within and across data sources $

David Milward, Matt Garber, Nicolas Joannin, and Qais Hatim #

Introduction

Why use ontologies for normalization of unstructured data:
- ✓ For better statistical analysis
- ✓ For input to machine learning models
- ✓ For linking of information within and across documents and data sources

This project concentrated on diseases and ingredients:
- ✓ UNII terminology, which associates ingredient terms with UNII codes,
  - Translated this into the ANSI thesaurus standard
  - It was incorporated into the Linguamatics NLP platform, i2e, and incorporated into FDA Drug Labels #
  - Used by FDA staff, e.g. to compare overdosage sections across drugs with the same ingredients
- ✓ Disease coding was investigated with state-of-the-art methods based on BERT (Bidirectional Encoder Representations from Transformers).

[PT] PHENOBARBITAL SODIUM	PHENOBARBITAL SODIUM	#Docs	#Hits	Hit
PHENOBARBITAL SODIUM	phenobarbital	272	2	phenytoin, sodium valproate, phenobarbital
	Phenobarbital	55	1	Phenobarbital, valproate sodium, valproic ...
	phenobarbital sodium	48	10	Each mL contains phenobarbital sodium either 65 mg or 130 ...
	Phenobarbital sodium	48	2	Phenobarbital sodium may be administered intramuscularly or ...
	Phenobarbital Sodium	4	19	Phenobarbital Sodium Injection, USP CIV FOR ...
	2,4,6(1H,3H,5H)-Pyrimidinetrione,5-ethyl-5-phenyl-, monosodium salt	2	1	Chemically, phenobarbital sodium is 2,4,6(1H,3H,5H)-Pyrimidinetrione,5-ethyl-5-phenyl-, monosodium salt and has the following structural ...
	PHENOBARBITAL SODIUM	1	2	PHENOBARBITAL SODIUM
	PHENOBARBITAL Sodium	1	2	CIV PHENOBARBITAL Sodium Injection, USP

Concept normalization approaches

Two typical approaches:
- Matching terms directly
  - Using terminologies directly on the data source.
  - Synonyms for each concept are matched to the text, either with an exact match or with some variation (e.g. morphological variants).
  - Ambiguity handled through confidence value using a disambiguation algorithm
- Matching via named entities
  - Using Named Entity Recognition (NER) algorithm to annotate a class of concepts e.g. chemicals
  - Terms found are then mapped to the individual chemicals

Machine learned approaches for Named Entity Recognition

- Require representative, large-scale, annotated data
- Used in cases where there is no existing comprehensive terminology

In recent years **BERT models** have been particularly prominent and provide excellent results for recognizing e.g. organizations or person names.
These are particularly problematic for a terminology approach since they are forever changing.

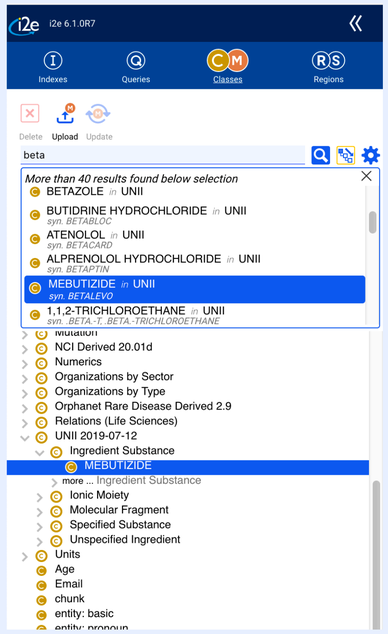
A terminology such as **UNII** does not fit well with standard named entities such as chemicals. The notion of an ingredient is broader, encompassing concepts such as *air*, *cork*, *blood* and *EGFR*. To train a ML model to recognize ingredients would require text specifically annotated for the ingredient class. A faster first step is to use a terminology directly. In future this could be used to semi-automatically annotate text to train an ML model.

Using a terminology for text mining

This typically involves 4 stages:
- ✓ Conversion of format
  - Terminologies come in many formats including TSV, bespoke formats such as OBO and OWL, or even Excel
- ✓ Determining the best options for matching the text e.g.:
  - morphological variants
  - fuzzy matching
  - OCR correction
- ✓ Checking for issues with noise and recall
- ✓ Refining to reduce noise and increase recall

Disease normalization

- Due to PHI issues, training data for healthcare is limited
  - Requires methods that can work well on unseen data, and are not overfitted to the data they were trained on
  - Previous work contrasted BioBERT to Clinical BERT, and suggested that BioBERT works well on new data even without retraining
  - To check this we annotated a new data set based on medical transcriptions



In this work we wanted to compare # the terminology approach with a ML # approach. This unseen data set was # used to evaluate both methods.

ML Approach

- We used a BioBERT disease NER model as implemented by BERN (Biomedical named Entity Recognition and multi-type Normalization)
- BioBERT is an offshoot of BERT, a state-of-the-art pre-trained bidirectional language model that can be fine-tuned for various NLP tasks
- BioBERT has been further pre-trained on a large corpus of PubMed and PubMed Central documents and has been shown to outperform BERT on tasks in the biomedical domain
- BERN is a tool that implements several different trained BioBERT models that have been fine-tuned for NER on several classes on entities, such as diseases, genes, and species.

Terminology Approach

This is similar to the approach taken for UNII described above, although the Diseases terminology does incorporate information from both MeSH and NCI to create a larger number of synonyms than MeSH would provide on its own. #

Evaluation

Currently our evaluation is at the class level, i.e. recognition of Diseases. Ideally we would provide further annotation of this data set to also evaluate at the individual concept level.
Results of the evaluation even at the recognition level were as follows.

Method	F-Score
BioBERT NER	73
BioBERT NER with Noise Reduction	74
Diseases Terminology	80
Combination of BioBERT and Diseases	82
Diseases Terminology augmented with terms found by BioBERT	84

Figures on this medical transcriptions dataset can be contrasted with the NCBI corpus (used to train BioBERT NER) where the Diseases terminology had the same figure of 84, but BioBERT achieved 91%.

Conclusions

The UNII ingredients ontology has now been incorporated within the processing of Drug Labels, and has been used, for example, to compare overdosage sections across drugs with the same ingredients.

Although use of BERN on its own did not improve results for diseases on this unseen data, comparison between BERN and the terminology results was useful, providing new synonyms that have been added to the terminology to improve results further.

Next steps

In this project we looked at Disease normalization using MeSH since this was already supported by BERN. The next step is to use a similar approach for MedDRA. We would also like to extend our evaluation to the concept normalization level, not just the recognition level. Given the large differences seen between datasets that were used for training vs. unseen datasets we also need to explore the level of training required to get similar performance.