Medical Named Entity Recognition on Twitter Data

March 05, 2016

Submitted By:-

Juhi Tandon - 201225032
Kalpish Singhal - 201505513
Chinmay Bapna - 201302182

Description (Problem Statement):

Medical Entity Recognition is a crucial step towards efficient medical texts analysis.

The task of a Medical Name Entity Recognizer is twofold -

(i) identification of entity boundaries in the sentences.

(ii) entity categorization.

Our objective is to extend medical entity recognition for tweets.
Medical entities can be diseases, drugs, symptoms, etc. Previously, researchers in the field have used hand crafted features to identify medical entities in medical literature. It has been found that in contrast with semantic approaches which require rich domain-knowledge for rule or pattern construction, statistical approaches are more scalable.

Dataset:

We have a dataset of 1 year of tweets about 4 diseases and 32 drugs. A team of domain experts has annotated about 2000 tweets with entities (around 20 types: diseases, drugs, symptoms) and also relations (around 40 relation types: cures, causes, etc).

Applications:

1. To get feedback for different types of treatments available/applicable on certain diseases (for which research is still ongoing) or mutation in disease causing microbes making them resistant to previously working drugs.
2. For getting information on side-effects of new drugs.
3. For getting information on symptoms of newly evolving diseases.

Challenges:

1. Application of normal text-parsing rules not applicable.
   Examples:
   - Multiple Sclerosis is a disease name but **Multiple** could be mistaken for a common terminology here
   - AIDS - Difficult to grasp when use of full form
2. Tweets are informal, noisy with linguistic errors and idiosyncratic style which degrades the performance of NLP tools on them.
3. Learning distributed representations for medical tweets.
4. Use of Non standard medical terminology
5. Entities can range from being a token to a sentence. Thus detection and delimitation of phrasal information referring to medical entities in textual corpora is a task.
6. Tokenisation seems difficult - can't remove numbers, special characters, etc
7. Spell correction and normalisation is difficult
8. Entity linking for exploiting semantic features from ontologies (UMLS, MetaMap).

Second Deliverable:

- Feature Identification and Listing.
- Add semantics to tweets by automatically identifying concepts that are semantically related to it .
- Conduct an empirical analysis of Medical named entity recognition and disambiguation.
- Application of Machine Learning algorithms and Feature Engineering in order to obtain state of the art results.
- Explore the hybrid approaches that aim to combine the advantages of semantic and statistical approaches

Third Deliverable:

- Accuracy enhancement.
- Reducing the need for manual inspection and selection.
- Along with improvements in Feature Engineering, also trying some Deep Learning Algorithms such as RNNs and LSTMs which can give better results.

Tools:

1. CMU's NER toolkit for twitter http://www.cs.cmu.edu/~ark/TweetNLP/
2. NLTK tools
3. CRF++ (Conditional Random Fields)
4. Maven
5. MetaMap and UMLS

References:

1. http://www.aclweb.org/anthology/D11-1141
2. http://www.springer.com/cda/content/document/cda_downloaddocument/9783319227405-c2.pdf?SGWID=0-0-45-1524297-p177643049
3. http://pnrsolution.org/Datacenter/Vol3/Issue6/66.pdf
4. http://www.aclweb.org/anthology/W11-0207
5. http://dl.acm.org/citation.cfm?id=2002911
6. https://github.com/aritter/twitter_nlp