VIRGINIA COMMONWEALTH UNIVERSITY

VCU School of Engineering
VIRGINIA COMMONWEALTH UNIVERSITY

Automatically extracting crucial nanomedical characteristics

Entity Extraction for Nanoinformatics

Gabrielle Jones, Nastassja Lewinski, PhD, Bridget McInnes, PhD

Introduction

- Natural Language Processing can be used for entity extraction to advance the progress of nanoinformatics, which combines the fields of computer science, life sciences, and chemical engineering.
- Entity extraction will aid in the analysis of new connections between different nanomedicines and their characteristics.
- Evaluation of different existing entity extraction systems will give insight into what algorithms can be utilized specifically for nanoinformatics applications.

Entity Extraction Systems

This research focuses on evaluating the existing entity extraction systems :

- Apache's OpenNLP
- Stanford NLP
- Banner
- Abner

Data

Manual Annotated Data:

52 FDA labels were manually read and instances that contained relevant characteristics were annotated.

Seed Data:

Primary literature was automatically annotated based off of seed patterns extracted from the manually annotated data.

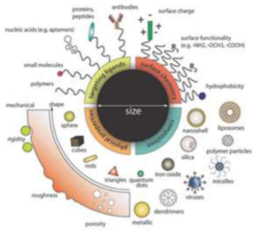


Figure 1: Nanoparticle characteristics

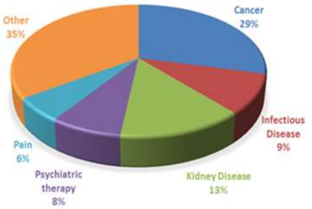
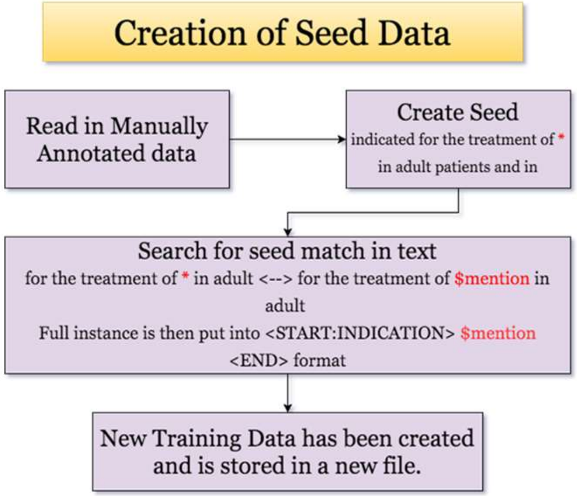


Figure 2. Disease classes treated by the 52 FDA approved nanomedicines.

How Does it Work?



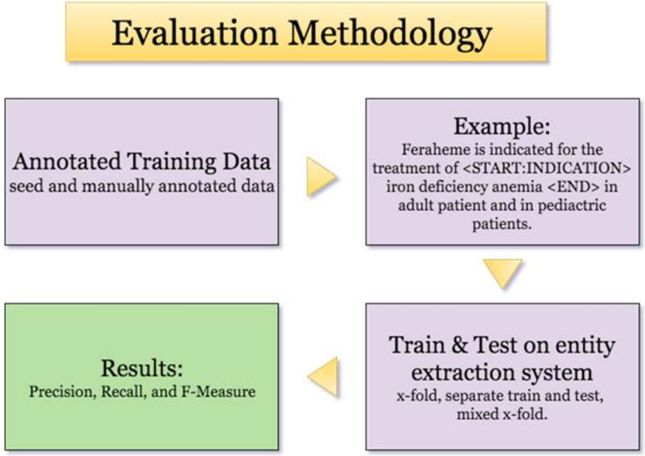
Creation of Seed Data

Read in Manually Annotated data

Create Seed
indicated for the treatment of * in adult patients and in

Search for seed match in text
for the treatment of * in adult <--> for the treatment of $mention in adult
Full instance is then put into <START:INDICATION> $mention <END> format

New Training Data has been created and is stored in a new file.



Evaluation Methodology

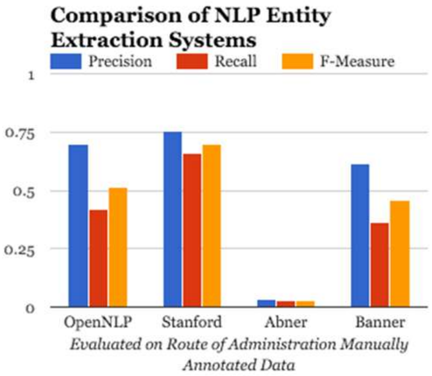
Annotated Training Data
seed and manually annotated data

Example:
Feraheme is indicated for the treatment of <START:INDICATION> iron deficiency anemia <END> in adult patient and in pediactric patients.

Train & Test on entity extraction system
x-fold, separate train and test, mixed x-fold.

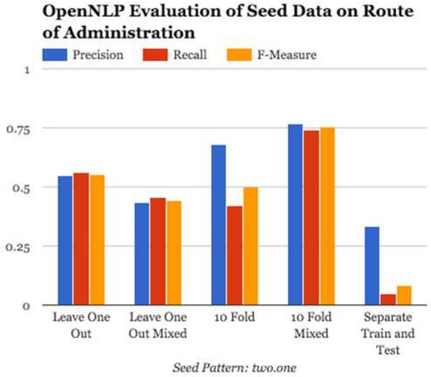
Results:
Precision, Recall, and F-Measure

Results



Comparison of NLP Entity Extraction Systems

Evaluated on Route of Administration Manually Annotated Data



OpenNLP Evaluation of Seed Data on Route of Administration

Seed Pattern: two.one

Conclusions

- Inclusion of automatically generated seed data increases the accuracy of the system.
- Context in FDA labels is consistent with primary literature.

Future Work

- Replace drug specific patterns with a generic marker.
- Options for different pattern requirements for creating seed data.