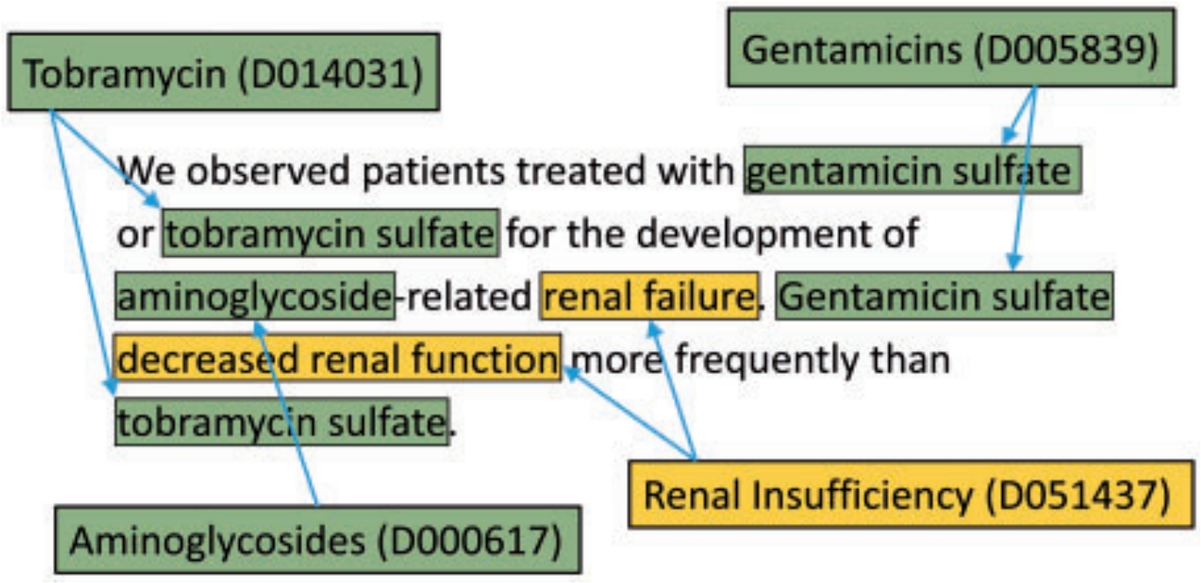
Biomedical Concept Normalization

Xuan Wang (xwang174@illinois.edu)
Fei Ling (fling2@illinois.edu)

Introduction

Biomedical concept normalization is a task to link the biomedical entity mentions in the text to a concept identifier in a controlled vocabulary.



Leaman et al., *Bioinformatics* 2016

Related work

- **Baselines:**
  - tmChem [1]:
    - Named entity recognition: combine of two CRF models.
    - Concept normalization: lexicon (MeSH and ChEBI) look-up.
  - DNorm [2]:
    - Named entity recognition: CRF based on rich features.
    - Concept normalization: pairwise learning to rank.
- **State-of-the-art:**
  - TaggerOne [3]: joint entity recognition and concept normalization with semi-Markov model.

Methods

- Dictionary look-up.
- Pairwise learning to rank.
  - Margin ranking perceptron: *m* is the representation of entity mention in text, n^+ is the true label of *m* in the concept ontology, n^- are all the labels in the concept ontology except for n^+.

$$W = \underset{W}{\operatorname{argmin}} \sum_{m} \sum_{n^+} \sum_{n^-} max\left(0, 1 - m^T W n^+ + m^T W n^-\right)$$

- Dimension reduction
- Negative sampling

Results

Overall statistics of the BioCreative V CDR corpus

Task Dataset	Articles	Chemical		Disease	
		Mention	ID	Mention	ID
Training	500	5,203	1,467	4,182	1,965
Development	500	5,347	1,507	4,244	1,865
Test	500	5,385	1,435	4,424	1,988

Micro-average scores of the two methods on the CDR corpus

	Precision	Recall	F1
Dictionary look-up	0.39	0.39	0.39
Pairwise learning to rank	-	-	-

Demo

- Demo with dictionary look-up.

Discussion

- Why the current implementation of pairwise learning to rank failed?
  - Current candidate label space is too large. It contains 270,000 concept identifiers in the MeSH ontology.
    - The next step is to use those most related concept identifiers (e.g., chemicals and diseases) as candidate labels for ranking.
- The representation of entity mentions can be future improved.
  - Current features are tf-idf vectors of the entity mentions.
  - The next step is to use the representation learned from a neural network model (e.g., Bi-LSTM) as the features for ranking.

References

1. Leaman, Robert, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization." Journal of cheminformatics 7, no. 1 (2015): S3.

2. Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu. "DNorm: disease name normalization with pairwise learning to rank." Bioinformatics 29, no. 22 (2013): 2909-2917.

3. Leaman, Robert, and Zhiyong Lu. "TaggerOne: joint named entity recognition and normalization with semi-Markov Models." Bioinformatics 32, no. 18 (2016): 2839-2846.