Named Entity Recognition in Bio Medical Text

CSE 5539: Natural Language Question Answering

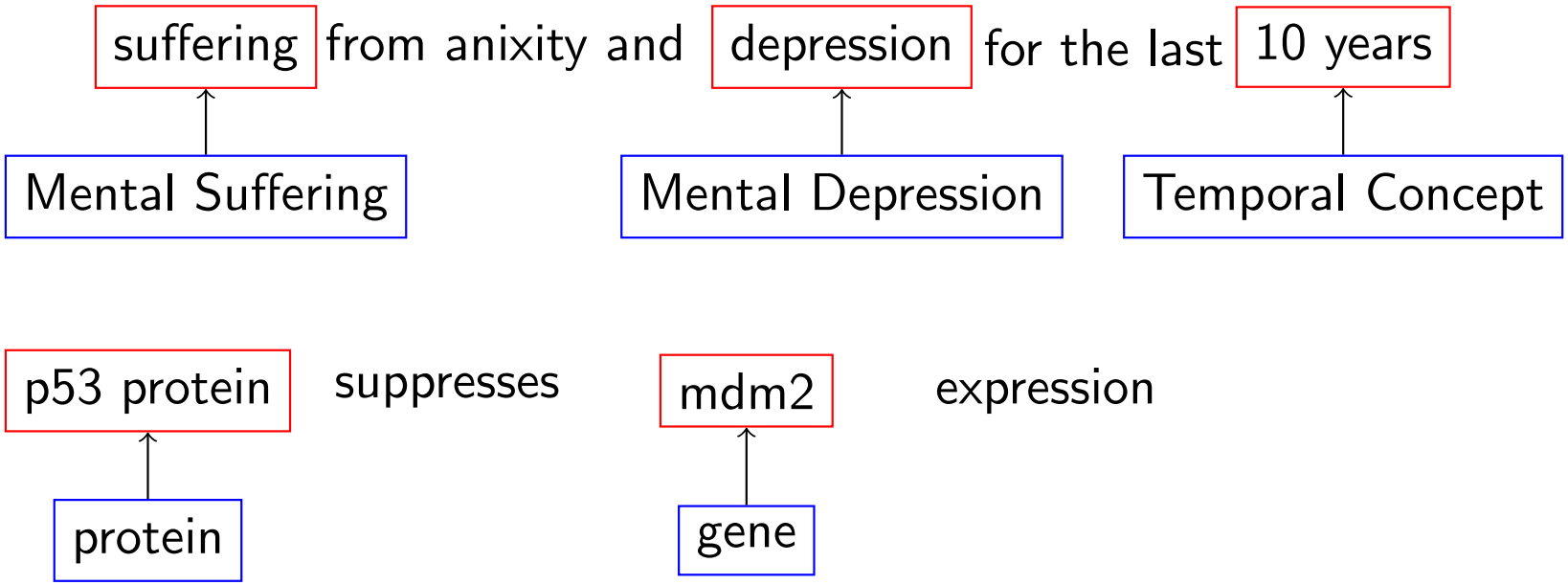
Dushyanta Dhyani , Diwen Hu

Department of Computer Science
The Ohio State University

Outline

Problem Definition & Motivation

- **What do we want?**



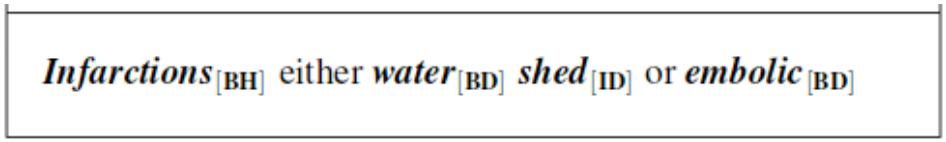
- **Motivation** -
  - Biomedical literature is increasing in volume.
  - Online Discussion Forums specific to Medical domain are being formed.
  - Thus automated information extraction and Knowledge Base Construction has become essential.
  - Named Entity Recognition forms one of the fundamental steps in most of the above procedures.

Datasets and Tools

- Datasets
  - Unified Medical Language System (UMLS) [1] Semantic Network
  - CHQA Named Entity Dataset [2]
- Existing Tools - Baselines
  - Metamap [3] - Maps biomedical text to UMLS entities
    - One of the largest medical KB.
    - Uses Locality sensitive hashing for recognizing entities.
  - BANNER [4] - Uses Conditional Random Fields with various Syntactic Features for tagging Named Entities.
  - Various other HMM based tools [5] [6]

Challenges

- Metamap uses simple string similarity and thus produces low scores.
- BANNER, a CRF based tool suffers from the implicit inability of sequence models to identify discontinuous entities.
- To enable CRF's to identify discontinous entities,an extended form of the **BIO** tagging which includes **B,I,O,BD,ID,BH,IH** is proposed in [7]
- Example

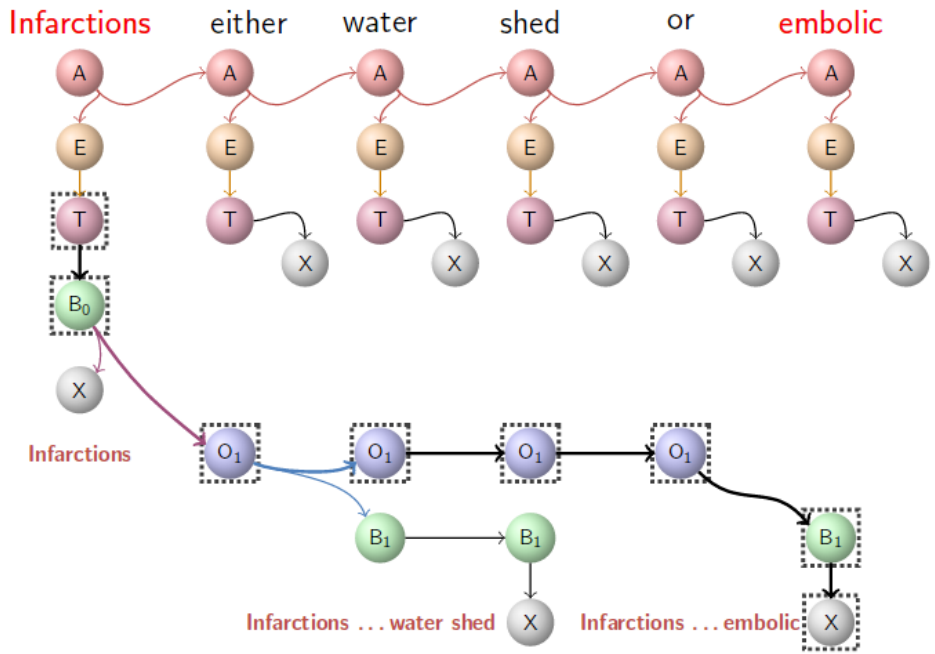


[8]

- Even if all tokens are correctly annotated, determining all the corresponding entities is difficult.
- The annotation scheme is not directly applicable to the CHQA dataset since entities in this dataset do not necessarily overlap at the beginning

Recognizing Discontinuous Entities

- We are currently trying to adopt the idea from [8] (referred onwards as statnlp tool) which models the task of Discontinuous and overlapping named entities as a hypergraph prediction problem where various nodes are :



A - All entities that begin with current or future word.

E - All entities that begin with the current word.

T - Entities of a specific type t that begin with the current word.

B - The word is part of the component of a given entity type.

O - Given word is between two components of a given entity

Recognizing Discontinuous Entities

- Each entity is represented by a unique subgraph.
- Given this hypergraph based model and training for maximizing the conditional log-likelihood of the data, the task now becomes determing the most likely sequence (MLS) or the Highest-Scoring subgraph using the Viterbi Algorithm.
- Challenges - The tool requires the knowledge of UMLS entity types. However, the CHQA dataset is annotated with newly created wrapper entities on top of UMLS entity types. e.g. UMLS Types -(Disease, Syndrome, Neoplastic process) have been combined into **PROBLEMS** in CHQA.

Current Progress

- Performed initial analysis of Metamap on the CHQA dataset.
- Performance of Metamap is low - 13%
- Tested Banner on non-overlapping and non-continuous entities from the CHQA dataset and on a 5-fold cross validation split achieved the following best scores -
  - **Precision** - 80.58%
  - **Recall** - 67%
  - **F1-Score** - 73.17%
- Currently working on empirical evaluation of statnlp tool on CHQA dataset.

References I


Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray.
The unified medical language system.
Methods of information in medicine, 32(4):281–291, 1993.


Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Kirk Roberts, Laritza Rodriguez, Sonya E Shooshan, and Dina Demner-Fushman.
Annotating named entities in consumer health questions.


Alan R Aronson.
Effective mapping of biomedical text to the umls metathesaurus: the metamap program.
In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

References II

Robert Leaman, Graciela Gonzalez, et al.
Banner: an executable survey of advances in biomedical named entity recognition.
In *Pacific symposium on biocomputing*, volume 13, pages 652–663, 2008.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan.
Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain.
In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 49–56. Association for Computational Linguistics, 2003.

References III


Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii.
Extracting the names of genes and gene products with a hidden markov model.
In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 201–207. Association for Computational Linguistics, 2000.


Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu.
Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features.
BMC medical informatics and decision making, 13(1):1, 2013.


Aldrian Obaja Muis and Wei Lu.
Learning to recognize discontiguous entities.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP16)*, 2016.