Named Entity Recognition in Bio-medical Text
Project Report

Dushyanta Dhyani
dhyani.2@osu.edu

Under the supervision of
Prof. Huan Sun

1 Introduction

Named Entity Recognition (NER) in Natural Language Processing is considered to be a problem that is "mostly" solved. However, when it comes to Bio-medical NER, the problem is still largely unsolved, due to the following reasons:

- Long Bio-medical names. e.g. **Phenylephrine Hydro-chloride 10 MG Oral Tablet**
- Nested entity names e.g.$< PROTEIN >< DNA >$ **kappa3** $< /DNA >$ **binding factor** $< /PROTEIN >$
- Discontinuous Entity Names - e.g. **body** to turn **red**
- Irregular Naming Conventions. e.g. **NF-kappaB, NFKappaB, NF-kappa B**,etc.

Since NER is a fundamental building block to any downstream applications like Knowledge Base Construction or Domain Specific Question Answering, it requires the development of a system with high accuracy. [1] states that NER errors have the highest impact on the task of KB construction. Thus it becomes essential to investigate existing approaches for the task and formulate new methods. In this work we perform qualitative and quantitative evaluation of existing approaches and enlist future tasks to be performed.

2 Datasets

- **Unified Medical Language System [2] Semantic Network** - Determines and defines key terminologies, classifications, coding standards for bio-medical information systems
- **Consumer Health Question-Answers (CHQA) Named Entity Dataset [3]** - 1548 Consumer Health Questions submitted to National Library of Medicine (NLM) annotated with 15 broad categories. Contains about 15K entity annotations.

3 Tools/Techniques Evaluated

- **Metamap** [4] - Performs shallow string similarity techniques to map text strings to UMLS entities.
- **BANNER** [5] - A Conditional Random Field based tool with domain independent inbuilt feature engineering techniques.
- **Statnlp**[1] - A recently published technique that models the problem of NER as a hypergraph prediction problem by using the Viterbi algorithm to predict the most likely sequence.

4 Evaluation Methodology

- **Metamap** - Since metamap does not require any form of training, the entire dataset was used as a test set for evaluation. Since metamap predicts multiple results as possible candidate answers, we select prediction as correct if any of the candidate answer is correct. The scoring was done for entity level matching i.e. correct if entire entity is predicted correctly.
- **BANNER** - Since CRF's are suited for sequence prediction (which is the aim of this task), we intended to use the **B,I,O,BD,ID,BH, and IH** annotation scheme (an extension of the BIO scheme) as suggested in [6]. The **BH** and **IH** annotations are used for components of nested/shared entities while the **BD** and **ID** are used for components of discontinuous entities. However, since overlapping entities in the CHQA dataset do not necessarily have shared components in the beginning (H in BH and IH stands for head), using this scheme did not seem very intuitive. Moreover, as reported by [6], even after using this scheme, the model inherently has a high level of ambiguity. Thus, at present the evaluation has been done for non overlapping and non discontinuous entities.

[1] The tool does not have an official name but was release by a research group with the same name and is thus used here for naming

- **Statnlp** - 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. Thus no empirical evaluation of the tool is currently available.

5 Results

- **Metamap** - Metamap exhibited an accuracy of 13% on the entire CHQA dataset. A lot of the errors can be attributed to it's inability to identify correct entity spans in the first place (instead of incorrect annotations). Of the 15K entities in the dataset, it was also able to identify 676 entities correctly but classified them incorrectly.
- **BANNER** - On a 5-fold cross validation split, an F1-Score of 73.17% was achieved with a precision of 80.58% and Recall of 67% (a characteristic similar to Metamap as shown above). A lot of incorrect classifications made by BANNER were cases where the words were not biological names but rather common English phrases like *Shortness of Breath* or *yellow colored liquid*

6 Conclusion

In this study , we performed empirical evaluation of certain existing tools for Bio-NER on a newly released dataset. We also highlighted certain other methodologies that would be a part of future evaluation. We also identified certain common issues with the problem of BIO-NER which could be addressed in future work.

References

[1] Patrick Ernst, Amy Siu, and Gerhard Weikum. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC bioinformatics*, 16(1):1, 2015.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Aldrian Obaja Muis and Wei Lu. Learning to recognize discontiguous entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 75–84, Austin, Texas, November 2016. Association for Computational Linguistics.