MetaMapLite in Excel: Biomedical Named-Entity Recognition for Non-technical Users

Ravi Teja Bhupatiraju¹, Kin Wah Fung¹, Olivier Bodenreider¹
¹National Library of Medicine, Bethesda, MD, USA

Abstract
We developed an easy-to-use tool for non-technical biomedical researchers to conveniently conduct Named-Entity Recognition (NER) on biomedical text, in a familiar spreadsheet system.

Keywords
Natural Language Processing; UMLS

Introduction
In the past, Named Entity Recognition (NER) was an expert task performed by computer scientists. However, in the recent years, the application of NER has become more pervasive. Biomedical researchers, who may not be computer scientists or have strong technical skills, wish to apply NER methods and tools to help answer research questions.

MetaMap (https://metamap.nlm.nih.gov/) is one of the most popular tools for biomedical Named Entity Recognition (NER), more specifically for identifying terms from the Unified Medical Language System (UMLS) Metathesaurus in biomedical text. MetaMapLite is a Java reimplementation of the original MetaMap. Running these tools on biomedical text and parsing their output generally requires some programming skills, which places them out of reach for non-technical users. Our objective is to make biomedical NER tools easier to use by non-technical users.

Previously, the National Center for Biomedical Ontology (NCBO) provided such a tool for non-technical users, the BioAnnotator REST service, along with an Excel addin. However, this tool is no longer available.

Methods
After observing non-technical user issues with exploiting the software resources, we have created a simple system that allows a non-technical user to use MetaMapLite from an Excel spreadsheet such that interactive and batch processing operations may be performed from a familiar spreadsheet environment.

Our system does not depend on network or external resources. Instead, a zero-configuration backend server provides an HTTP service that a spreadsheet macro-implemented HTTP client consumes to provide a single, flexible, easy-to-use Excel function to perform named entity recognition. The function features output field selection and semantic type restriction. The backend serves up a self-documenting spreadsheet template for the users to get started. The baseline spreadsheet functionality allows for automatically updating NER results as users edit entries while conducting interactive exploration and for conveniently dragging fill handles to apply the function to batches of natural text inputs. The function may be used in macros for further automation solutions.

Results
Figure 1 illustrates a typical use case for our tool. Users copy biomedical text in one column (A) and use the `mmlite` function in another column (B) to identify UMLS concepts from the text in column (A).

Users found the software easy to install and use. Response times were quick, at about 30ms per request on a Xeon E5-1620 v3 3.5 GHz with 16 GB RAM.

From a technical perspective, the backend component does not have any special hardware requirements and can run anywhere a Java Virtual Machine (JVM) is available. The user package is a 465 MB Windows installer bundling JVM 1.8, Apache Tomcat 8.5.4 and MetaMapLite 3.0. The client was implemented in Visual Basic for Applications with WinHTTP and therefore works only on MS Windows for now.

Conclusion
Biomedical NER can be provided in a convenient package for non-technical users. In the future, we will explore extending the spreadsheet client to support all major desktop platforms, covering the original MetaMap and providing a publicly hosted REST API to remove need for backend installation.

Address for correspondence
raviteja.bhupatiraju@nih.gov

Figure 1-Usage scenario demonstrating the use of the mmlite function in conjunction with fill handles in quickly applying NER. Field selection specified preferred name and semantic type restriction was used to only identify pharmaceutical substances.