MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-1252

MetaMap Lite in Excel: Biomedical Named-Entity Recognition for Non-Technical Users

Ravi Teja Bhupatiraju, Kin Wah Fung, Olivier Bodenreider

National Library of Medicine, Bethesda, MD, USA

Abstract

We developed an easy-to-use tool for non-technical biomedical researchers to conduct Named-Entity Recognition (NER) on biomedical text, in a familiar spreadsheet environment. The system is a simple, offline, easy to install, end-user front-end to the new MetaMap Lite. Early adopters found it to be a quick starting-point to incorporate NER in their investigations.

Keywords:

Natural language processing; Unified medical language system.

Introduction

The application of Named Entity Recognition (NER) has become pervasive. Biomedical researchers, who may not have strong computer skills, often wish to apply NER methods and tools to extract information from text.

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. *MetaMap Lite* is a recent 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.

Methods

We developed an easy-to-use tool for non-technical biomedical researchers to use *MetaMap Lite* on biomedical text, in a familiar spreadsheet environment, supporting interactive and batch processing operations.

Our system does not depend on network or external resources. Instead, a zero-configuration backend server provides an HTTP service that a spreadsheet function consumes to perform named entity recognition. The function supports output field selection (e.g, "pref,stype" returns the preferred name and semantic type, along with the UMLS concept unique identifier, or CUI). Matched text and source vocabulary may also be requested. By default, the system only returns the UMLS CUI and the preferred name). Semantic type restriction may also be specified (e.g., "phsu,antb" returns only those terms that have been categorized as pharmaceutical substances or antibiotics).

The backend server serves a self-documenting spreadsheet template for users to get started. It supports automatic update of NER results as users edit entries, and batch processing by dragging fill handles to apply the function to rows of natural text inputs. The function may be combined with other functions for further automation.

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).

From a technical perspective, the *backend* can run anywhere a Java Virtual Machine (JVM) is available. The Windows installer for the software package contains all the necessary software components for running the *mmlite* function in Excel.

On informal inquiry, 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.

Acknowledgements

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Address for correspondence

Olivier Bodenreider: olivier@nlm.nih.gov

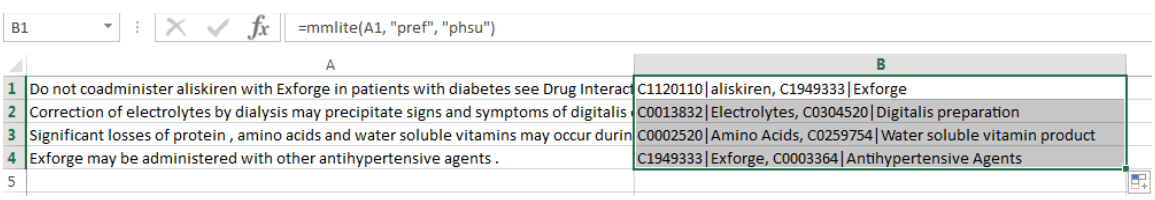

Figure 1 - Example of use of the mmlite function in conjunction with fill handles in quickly applying NER