Exploring Multi-terminology Indexing of Discharge Summaries

Suzanne PEREIRA¹,²,³, Philippe MASSARI¹, Michel JOUBERT², Elisabeth SERROT³ and Stefan DARMONI¹

¹ CISMeF & LITIS EA 4108, Institute of Biomedical Research (Rouen, France)
² LERTIM, Marseille Medical University (Marseille, France)
³ VIDAL Company (Issy les Moulineaux, France)

Keywords: Terminology-vocabulary; Coding-coding systems; indexing

The coding of medical diseases and procedures involves the translation of medical words from various reports into codes that accurately reflect the state of health and the care that patients have received. Coding is a complex process that can be performed manually by an educated and trained individual respectful to the coding standards rules. Coders can be helped by automatic tools to capture and search for ICD10 codes but also to extract ICD10 codes from medical texts. The goal of this study is to describe and to evaluate F-MTI (the French Multi-terminology Indexer tool) to improve ICD10 codes assignment to French hospital discharge summaries.

F-MTI works as follows: 1) all the headings and their content (“medical history”, “conclusions...”) are separated and all the sentences are extracted from discharge summaries; 2) each sentence is normalized (accents and stop words are removed, all words are switched to lower case...) to form a “bag of words”. This bag is matched against all the MeSH, SNOMED, ICD10, and CCAM terms that have been previously processed in the same way. All the terms retrieved are projected in ICD10. The candidate set of indexing terms produced by F-MTI for the discharge summary is the result of the union of the terms extracted directly (from the matching between the bag of words of the sentence and those of all the terms of the different terminologies) and more original the terms extracted indirectly (from the projection in ICD10).

The contribution of F-MTI was measured on a sample of 100 discharge summaries from pneumology et and cardiology services. It performed a precision of 2.6% of and a recall of 38.0% compared to a manual medico-economical indexing and a precision of 3.7% and a recall of 32.9% compared to a manual descriptive coding.

We also observed the results through different points of view restricting the codes to those related to the respective domain of cardiology and pneumology. To perform these restrictions to one medical specialty, we used metaterms previously developed for the MeSH thesaurus, and later adapted to the ICD10 codes. We also restricted these codes to the “disease or syndrome” and the “sign or symptom” codes. These annotations are semantic types linked to ICD10 codes that we extracted from the UMLS Metathesaurus.

Then F-MTI is close to a descriptive coding as it extracts all the terms it can found without applying any economical rules of coding. Further evaluation will be based on a corpus of at least 1,000 discharge summaries. Future work will address extending this work to other terminologies and documents indexing.