Knowledge Extraction of Thoracic Radiology Reports Using Statistical Natural Language Processing

Leandro Zerbinatti; Lincoln de Assis Moura Jr
izerbinatti@gmail.com

Introduction

This work promotes a study in health informatics technology which analyses reports of chest X-ray through statistical natural language processing methods for the purpose of supporting the interoperability between health systems. Zipf's constant was studied and it was determined that few words make up the majority of the reports and that most of the words do not have statistical significance.

The translation and comparison with existing standardized medical vocabulary with international terminology, called SNOMED-CT, was done based on the terms identified. The terms that had a complete and direct correlation with the translated terms were incorporated into the reference terms along with its class and the word identifier.

Material and Methods

The method applied in our work is depicted in Figure below and were synthetically were performed:

- The selection of the reference reports from the corpus. Reports were first selected in 1000 and later the study was repeated in 1000 to other reports, thus totaling 2000 reports;
- The application of statistical natural language processing, extracting the words, n-grams and phrases;
- The construction of the terms and referenced in Portuguese;
- The application of clinical reports labeler.

The clinical reports properly labeled with references to vocabulary are structured documents required for interoperability between systems based on ontologies described in OWL, the HL7 CDA and archetypes.

Results

Another 200 reports of chest x-rays were selected to perform the terms tagging experiment of with respect to the reference. The efficiency obtained, which is the percentage of labeling of the reports, was 45.55%.

Subsequently, articles, prepositions and pronouns were incorporated into the terms of reference under the linkage concept of class. It is important to note that these terms do not carry health knowledge to the text. Thus, the efficiency ratio was 73.23%, significantly increasing the efficiency obtained previously.

These word are not associated with a concept of SNOMED-CT. Its function is to assign or connect concepts within sentences. The improvement achieved is a direct result of the statistical representation that these words have in the experimental set.

Conclusion

Was confirmed attendance to Zipf's law for the corpus used. This law offers a few words that we use to communicate, and is based on the principle of least effort.

The labeling efficiency obtained was 45.55%. Are added to the 16 words including articles, prepositions and other words to link concepts, the efficiency of labeling was increased to 73.26%.

This increase in efficiency demonstrates that a few words have great statistical significance and are responsible for much of the composition of the corpus.

The patterns of representation of clinical information used to enable semantic interoperability were ontologies expressed in OWL, the HL7 / CDA and openEHR archetypes of the Foundation.

References