INFORMATION SYSTEMS IN HEALTH CARE:

facilitating for NLP and machine learning

Rosa Liliana Figueroa Iturrieta

Becaria Conicyt

Presentada en cumplimiento Parcial de los Requerimientos

para el grado de

Doctor en Ciencias

con mención Ingeniería Eléctrica

de la

Escuela de Graduados

de la

Universidad de Concepción,

Chile.

2012

Abstract

Biomedical texts are a rich source of information, once extracted, can suit several research purposes, such as understanding disease patterns, tracking epidemic outbreaks, and tracking drug side effects. Natural language processing (NLP) techniques in conjunction with machine learning (ML) algorithms are being used by researchers to automatically extract biomedical information. However, NLP and ML algorithms must face two challenges when processing biomedical texts: the ambiguity in the word sense and the annotation of a large training corpus. Automatic word sense disambiguation has become an essential task for improving the accuracy of NLP algorithms, which still requires large quantities of manually annotated training data as well as large vocabularies that do not always cover all the senses of a word.

In this thesis the issues of ambiguity and data annotation in NLP have been addressed by means of three methods. First, I proposed a method that reduces the need of word sense disambiguation without compromising NLP performance has been developed. The method tailors a comprehensive vocabulary system for a biomedical sub-domain, e.g. clinical reports, by detecting unused senses in a sub-domain and removing them from the vocabulary thereby facilitating NLP. Specifically, word sense detection is achieved by comparing the relational neighborhood of a word in the vocabulary with the semantic neighborhood of the word in the sub-domain.

The second method analyzes the effectiveness of applying sample selection techniques, such as active learning, to improve clinical text classification. The analysis conducted attempts to find out whether active learning methods are useful means to reduce the dimension of large clinical data sets that are used for the training purposes of ML. The key idea developed as a consequence of the analysis is to improve the learning process by selecting the most informative cases among all the cases available in the data sets. By doing so, the size of the required training samples is indeed reduced without compromising the performance of the learning process, and more importantly, reducing the costs associated with the manual annotation of documents.

The third method consists in modeling and fitting learning curves for predicting the performance of classifiers. Specifically, in this thesis the learning curves of the classifiers are modeled as nonlinear functions of the sample size. The applicability of the models is twofold: it can assist researchers to decide how much data should be annotated in order to achieve a desired performance level and also it can estimate the performance achieved by a given classifier when adding more annotated data to their training set.

In order to systematically assess the performance of the methods developed in this thesis work, theoretical models and algorithms have been developed; data analysis and evaluation have been carried out using biomedical texts. This thesis also introduces a new tool for annotation of clinical reports, which has been built based on the findings of active learning algorithms.

In terms of performance, the three methods developed have shown improvements when compared to their respective control methods. In the first method, unused sense detection showed an improvement of 10% in terms of area under the receiver operation curve (AUC) when compared to the assigned control method (AUC for unused detection were between 72% and 87%).

The application of active learning algorithms to medical text classification has shown that some active learning variations can effectively reduce the size of the training

set. Since less training examples are needed, active learning methods can reduce the burden of annotation and speed up the development of NLP methods. For example on the smoking set (SNS2 dataset) to reach an accuracy of 90% some active learning methods required one third fewer training examples than the traditional passive selection.

The third method provides as a result a weighted model and a fitting algorithm which can be used to predict classifier performance. For several representatives cases of learning curves, the provided weighted model showed to be more accurate on the predictions of require sample size than the unweighted control method ($p<0.05$).

Each of the methods I have developed, tested, and evaluated in this thesis are meant to facilitate NLP processing. I have treated the sense detection and training sample selection as separate problems.