Identifying human phenotype terms in text using a machine learning approach

Manuel Lobo[1], André Lamúrias[1], Francisco M. Couto[1]

[1]LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

The purpose of this work is to provide an automatic entity detection system for the identification of human phenotype entities in unstructured text. This system is being developed as an integrated part of IBEnt (https://github.com/AndreLamurias/IBEnt) [1].

IBEnt is Python-based Framework for identifying biomedical entities that relies on the use of Stanford's Natural Language Processing tool – StanfordCoreNLP [2] - which contains tools such as StanfordNer [3] (Stanford Named Entity Recognizer) and Part-of-Speech tagger.

With these tools, IBEnt is able to create different types of classifiers that allow the system to recognize entities for different purposes. This system is in constant development, making available different classifiers for chemical entities, protein entities and, in this case, phenotypical entities.

For this work, machine learning techniques are going to be applied to try to improve the quality of the recognition. One of these techniques is Brown clustering [4] that allows the creation of clusters that group words together, according to a statistical analysis.

To test the performance of the system, we are using Bio-Lark's Gold Standardized Corpora as well as their provided Test Suites created to benchmark human phenotype classification system [5].

As an additional point for this work, a secondary objective is to study the effect of automatic translation of terms to understand the amount of information that is lost during an automatic translation (translation to from English to a different language and back to English), to understand if it is possible to apply this kind of translation on a regular basis.

References

[1] Lamurias A, Ferreira D. J, Couto M. F. *Improving chemical entity recognition through h-index based semantic similarity*, 2015, J Cheminform, (Suppl 1 Text mining for chemistry and the CHEMDNER track):S13.

[2] Manning, Christopher D., Surdeanu M, Bauer J, Finkel J, Bethard J. S and McClosky D. *The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55-60

[3] Finkel R. J, Grenager T, and Manning C. *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 2005, pp. 363-370.

[4] Brown P. F, deSouza P. V, Mercer R. L, et al. *Class-based n-gram models of natural language*, 1992, Comput. Linguist. 18, 467–479

[5] Groza T, Köhler S, Doelken S, et al. *Automatic concept recognition using the human phenotype ontology reference and test suite corpora*, 2015, Database (Oxford).