Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220610

Automatic Prediction of Semantic Labels for French Medical Terms

Thierry HAMON[a,1] and Natalia GRABAR[b]
[a] *Université Paris-Saclay, CNRS, LISN, F-91400, Orsay, France*
Université Sorbonne Paris Nord, F-93430, Villetaneuse, France
[b] *CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France*
hamon@limsi.fr, natalia.grabar@univ-lille.fr

Abstract. We address the problem of semantic labeling of terms in two French medical corpora with the subset of the UMLS. We perform two experiments relying on the structure of words and terms, and on their context: 1) the semantic label of already identified terms is predicted; 2) the terms are detected in raw texts and their semantic label is predicted. Our results show over 0.90 F-measure.

Keywords. Semantic labeling, terminology, NLP, Machine Learning, French

1. Introduction

Semantic labeling of terms consists of assigning semantic type (e.g. disorders, procedures, medication, chemical components, anatomy, signs and symptoms) to a term which is given or identified in the documents. Recent initiatives motivated the research on semantic labeling, mainly on English medical texts for concept normalization [1,2], but also French medical texts [3]. We aim to predict semantic labels of medical terms, with or without the detection of their boundaries within texts. We rely on two corpora which are part of the CLEAR corpus [4]: the French Wikipedia (1,324 articles, 3M words) and Cochrane database (7,678 abstracts, 4.5M words). The 238,983 French terms of the UMLS [5] associated with one of the 15 semantic groups [6] are projected on the corpora. The annotations are used as reference data (respectively, 58,213 and 123,880 recognized terms).

2. Methods and Results

On the basis of the reference data, we perform two experiments. The word or term features are related to their structure (the inflectional form, prefixes and suffixes with 1 to 3 characters, presence of uppercased and lowcased characters, and presence of special characters and numbers) and to their context (inflectional forms, lemmas and POS-tags within the 5-word windows on the left and on the right. We evaluate the results with 10-

[1] Corresponding Author, Thierry Hamon, LISN, Campus universitaire bât 507, Rue du Belvedère, F - 91405 Orsay cedex, France; E-mail: hamon@limsi.fr.
This work was partly funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01.

fold cross-validation and standard macro-measures at the level of semantic groups: Precision, Recall, and F-measure.

In the first experiment, the terms are already identified in texts. The task is to predict their semantic group. The terms are classified through the 15 semantic groups with several algorithms (Decision Trees [7], Random Forest [8], and SVM [9]). The SVM provides the best results: it outperforms Random Forest by 0.30, for instance, and gives balanced values of Precision and Recall. The average of the performance for all semantic groups with SVM is above 0.94 in both corpora. Cochrane abstracts get slightly better results than Wikipedia articles. Results indicate that it is quite easy to differentiate the 15 semantic groups among them on the basis of term structure and context. The second experiment consists in classifying each word according to 60 tags to predict the term boundaries and their semantic group with a BILOU representation. We used several algorithms (CRF [10], BiLSTM-CRF [11], Multilayer Perceptron (MLP) [12]). CRF outperforms BiLSTM-CRF by 0.30 and MLP by 0.40. The neural approaches are outperformed by CRF certainly because they may require larger datasets for training. The average of the CRF performance remains very high as well, with over 0.93 F-measure, while we observe that the size of classes is important. This experiment also permits to find out the most probable sequences of classes.

3. Conclusions and Future Work

We presented experiments on the semantic labeling of terms in French medical corpora through 15 semantic groups from the UMLS. The best results are obtained with CRF (over 0.90 F-measure) which identifies the boundaries of terms within documents, and predicts semantic groups of terms. In future work, we will enrich the reference dataset with adjectival forms of terms, use a BERT model for the semantic labeling and use these predictions for helping the automatic text simplification.

References

[1] Luo YF, Sun W, Rumshisky A. MCN: A comprehensive corpus for medical concept normalization. J Biomed Inform. 2019;92:103132.
[2] Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc. 2020;27(10):1529-37.
[3] Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform. 2021;114:103684.
[4] Grabar N, Cardon R. CLEAR – Simple Corpus for Medical French. In: Workshop on Automatic Text Adaption (ATA); 2018. p. 1-11.
[5] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281-91.
[6] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. In: Proceedings of Medinfo. vol. 10; 2001. p. 216-20.
[7] Quinlan J. C4.5 Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
[8] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
[9] Cortes C, Vapnik V. Support-Vector Networks. In: Machine Learning; 1995. p. 273-97.
[10] Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML'01. 2001. p. 282-9.
[11] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;7(8):1735-80.
[12] Rosenblatt F. The Perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review. 1958;65(6):386-408.