June 2021

Automatic Prediction of Semantic Labels for French Medical Terms

Thierry HAMON [a], Natalia GRABAR [b]

[a] *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 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. Semantic labeling of terms consists of assigning semantic label to a given term. In medical area, semantic labels are related to more or less large categories of notions, such as disorders, procedures, medication, chemical components, anatomy, phenotype, signs and symptoms. We address this task as classification problem. We work with data in French: two corpora and the French subset of the UMLS. We perform two experiments. In one experiment, the terms are already identified and the task is to predict their semantic label. In another experiment, we start with raw texts and have first to detect the terms within sentences and then to predict their semantic label. The features used are related to the structure of terms and to their context. Our results show over 0.90 F-measure for both experiments.

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

1. Introduction

Semantic labeling of terms consists of assigning semantic type to a given term. In medical area, semantic labels are related to more or less large categories of notions, such as disorders, procedures, medication, chemical components, anatomy, phenotype, signs and symptoms. Semantic labeling may be useful in several NLP tasks (information retrieval and extraction, decision making, machine translation, automatic simplification...) because it provides high-level information on semantic nature of terms. For instance, semantic labeling of terms may enrich medical information tracking system that supports decision making or quality assurance of medical treatment [1]. In information retrieval and extraction, semantic labeling helps to focus on precise notions within documents [2,3,4]. In machine translation, terms that belong to categories like medication, chemical components or genes, which are often absent from translation dictionaries, may not be translated but transposed to the target language instead. In automatic simplification, terms labeled as anatomy or phenotype can be considered as difficult to understand and become good candidates for the simplification [5].

Semantic labeling is close to concept normalization or terminology mapping, during which terms from texts must be linked to their normalized forms in the standard terminologies. This task usually starts with the identification of terms within documents, which may be considered as first step. Then follow concept normalization and prediction of semantic labels for terms, which may be done jointly or independently. Recent initiatives motivated the research on concept normalization on English data: creation of manually annotated corpus for medical concept normalization [6] and organization of the NLP task for concept normalization in clinical documents to the UMLS concepts as part of the N2C2 NLP challenge [7]. 33 teams participated in the task. Among the top 10 performing teams, different approaches have been used: dictionary look-up with editing distance and approximate matching, deep learning architecture, computing of cosine distance between vectors, information retrieval approaches. One work addressed concept normalization in French texts [8]. The terms are already provided to the system and the system has to predict the UMLS CUI for each terminological sequence. The authors encode terms with contextualized embeddings and classify them via cosine similarity and softmax. This system outperforms other existing systems on the Quaero FrenchMed Corpus showing between 0.743 and 0.851 F-measure.

The purpose of our work is to predict automatically the semantic labels for medical terms, with or without the automatic detection of their boundaries within documents. In what follows, we first describe the material used and methods proposed. We then present the results and discuss them.

2. Material and Methods

We rely on two types of data in French: textual corpora and the medical terminology UMLS [9]. Two corpora are exploited: French Wikipedia and Cochrane database, which are part of the CLEAR corpus [10]. Cochrane database proposes systematic reviews of scientific literature on different topics (prognosis, treatment, diagnosis...) within the framework of Evidence-Based Medicine. We exploit 7,678 abstracts containing almost 4.5M word occurrences. Wikipedia is an online collaborative encyclopedia and we exploit 1,324 articles related to medicine and totaling over 3M occurrences of words. The corpora are pre-processed syntactically with Treetagger [11] and FLEMM [12] to compute the part-of-speech (POS) tags and lemmas of words.

The UMLS is exploited in two ways. First, we use the set of 15 semantic groups indicated in the first two columns of Table 1. We also use the 238,983 French terms, which are projected on corpus documents. This projection is done on raw and lemmatized documents. During this process, we take into account the case of characters in order to better recognize the abbreviations. Hence, if a word has less than 4 characters, this word is supposed to be an abbreviation and the case of characters is respected. Words with only one character, although they may correspond to terms or abbreviations, are ignored because of their ambiguity. The projection of the UMLS terms on documents permits to create the silver standard which we use as reference data.

On the basis of these reference data, we propose to perform two experiments:

exp_1 The terms are already identified within texts and the task is to predict their semantic group. This task is addressed through classification. The features are related to the structure of terms (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). The classification is done with several algorithms (Decision Trees [13], Random Forest [14], and SVM [15]). The results are evaluated with 10-fold cross-validation.

exp_2 The terms are not identified and the task consists first into finding the terms and then into predicting their semantic group. The identification of terms is also addressed as classification problem: each word in corpora is to be assigned in one of the BILOU (Beginning of the term, In-word, Last word of the term, Out word outside the terms, Unique word corresponding to one-word terms) classes. We tested several algorithms (CRF [16], BiLSTM-CRF [17], Multilayer Perceptron (MLP) [18], implemented in python within CRFsuite and Keras libraries). The features used with CRF and MLP are related to inflected forms of words, the predicted semantic tags of previous words, and the features related to the structure of words as described in previous point, and each word is considered within a 5-word context on the left and on the right. With BiLSTM-CRF, we use only inflected forms of words. The corpus is segmented into training (80%) and test (20%) sets. We also do a 10-fold cross-validation on the whole corpus. At a second step, for each term identified, the semantic label is predicted with the method described in the previous point.

We evaluate the results with standard evaluation measures: Precision (are the results exact?), Recall (are the results complete?), and F-measure (harmonic mean of Precision and Recall). The values are computed in their macro version at the level of semantic groups, so that there is no or less bias because of the unbalance in the dataset.

Label	*Meaning*	*# recognized terms*		exp_1		exp_2	
		Wiki	*Coch.*	F_{Wiki}	F_{Coch}	F_{Wiki}	F_{Coch}
ACTI	Activities & Behaviors	2,612	8,510	0.948	0.990	0.987	0.998
ANAT	Anatomy	6,951	5,183	0.967	0.975	0.993	0.997
CHEM	Chemicals & Drugs	5,085	9,532	0.900	0.957	0.969	0.989
CONC	Concepts & Ideas	6,375	26,210	0.956	0.990	0.990	0.998
DEVI	Devices	263	787	0.871	0.920	0.948	0.995
DISO	Disorders	11,659	26,139	0.965	0.988	0.983	0.997
GENE	Genes & Molecula	203	8	0.977	0.985	0.852	0.250
GEOG	Geographic Areas	2,879	2,365	0.982	0.993	0.989	0.997
LIVB	Living Beings	7,140	14,652	0.964	0.991	0.987	0.998
OBJC	Objects	2,049	1,754	0.871	0.800	0.965	0.951
OCCU	Occupations	2,820	2,459	0.965	0.920	0.990	0.972
ORGA	Organizations	623	992	0.848	0.797	0.959	0.919
PHEN	Phenomena	1,088	1,120	0.902	0.898	0.975	0.969
PHYS	Physiology	4,671	6,303	0.935	0.966	0.985	0.993
PROC	Procedures	3,795	17,866	0.881	0.946	0.977	0.991
Tot/Avg		58,213	123,880	0.947	0.975	0.970	0.934

Table 1. Semantic groups from the UMLS, reference annotations, and F-measure obtained in two experiments: (1) prediction of semantic labels for terms with SVM, and (2) detection of terms with CRF and prediction of their semantic labels with SVM

3. Results and Discussion

Columns 3 and 4 in Table 1 indicate the number of term occurrences recognized within the two corpora (Wikipedia and Cochrane). We can observe that disorders, procedures, ideas and living beings are the most frequent semantic groups. Genes and molecula occur seldom in these corpora. We can also see that the number of terms from each semantic group is uneven: this reference dataset is unbalanced. A manual analysis of this dataset indicates that terms are recognized within documents fairly well. Some of the recognized terms may miss qualifiers. For instance, within the syntactic group *neurectomie présacrée* (*presacral neurectomy*), only *neurectomie* is recognized to be the UMLS term because *neurectomie présacrée* and *présacré* are also missing in the UMLS. In other cases, some adjectives like *pelvien* (*pelvic*) or *utérine* (*uterine*) could not be mapped with the corresponding nouns *pelvis* (*pelvis*) and *utérus* (*uterus*).

In the following columns of Table 1 we present the results for the two experiments (the results are indicated in terms of F-measure):

exp_1 Semantic labeling of terms, which are already recognized in the documents: for a given known sequence, we have to predict its semantic group among the 15 groups possible. Among the tested algorithms, the SVM provides the best results: it outperforms Random Forest by 0.30, for instance, and gives balanced values of Precision and Recall. We keep to the SVM results in what follows. The F-measure values are shown in columns 5 and 6 in Table 1. The average of the performance for all semantic groups is above 0.94 in both corpora. Cochrane abstracts get slightly better results than Wikipedia articles. These high values indicate that it is quite easy to differentiate the semantic groups among them on the basis of term structure and context. Even the group related to genes and moleculas, which is very small, is recognized well, certainly because of specific structure of the terms.

exp_2 Starting with raw texts, the system chains up two tasks. It has first to recognize the sequences that correspond to terms and then to predict their semantic groups. Among the tested algorithms, 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. In what follows, we present the CRF results. They are shown in columns 7 and 8 in Table 1. The average of the performance remains very high as well, with over 0.93 F-measure. We can see that the size of classes is important. Hence, for the Genes and Moleculas semantic group, which receives very few assignments in the Cochrane abstracts, the performance is very low. This class decreases the overall results for the 15 groups in this corpus. An interesting issue is that this experiment also permits to find out the most probable sequences of classes, such as *BL* for terms composed of two words like *crampes/B-DISO menstruelles/L-DISO* (*Dysmenorrhea*) or *voies/B-ANAT nerveuses/L-ANAT* (*Neural Pathways*), or *BI* and *IL* for terms composed of more than two words like *qualité/B-CONC de/I-CONC vie/L-CONC* (*Quality of life*) or *anti/B-CHEM inflammatoires/I-CHEM non/I-CHEM stéroïdiens/L-CHEM* (*Non-steroidal anti-inflammatory agent*).

4. Conclusion and Future Work

We propose to work on semantic labeling of terms, which purpose is to predict semantic groups for terms. We use 15 semantic groups from the UMLS, and propose two experiments according to whether the term boundaries are already given or not. We address this problem through classification and test several machine learning algorithms including deep learning. We get the best results with CRF when defining the boundaries of terms within documents, and with SVM when predicting semantic groups for terms. The results obtained reach over 0.90 F-measure.

Our future work may address several issues: (1) improvement of the reference dataset created by enriching it with additional terms such as adjectival forms of terms (*pelvis/pelvien*), (2) exploitation of these predictions for helping the automatic text simplification when detecting complex terms which should be simplified, (3) test of other approaches for the semantic labeling of terms.

References

[1] Jang H, Song SK, Myaeng SH. Semantic tagging for medical knowledge tracking. In: Conf Proc IEEE Eng Med Biol Soc; 2006. p. 6257-60.

[2] Mikkelsen G, Aasly J. Manual semantic tagging to improve access to information in narrative electronic medical records. Int J Med Inform. 2002;65(1):17-29.

[3] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17(5):514-8.

[4] Uzuner O, South BR, Shen S, DuVall SL. 2010 I2B2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552-6.

[5] Koptient A, Grabar N. Fine-grained text simplification in French: steps towards a better grammaticality. In: Proc of ISHIMR 2020; 2020. p. 1-10.

[6] Luo YF, Sun W, Rumshisky A. MCN: A comprehensive corpus for medical concept normalization. J Biomed Inform. 2019;92:103132.

[7] 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.

[8] Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform. 2021;114:103684.

[9] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281-91.

[10] Grabar N, Cardon R. CLEAR – Simple Corpus for Medical French. In: Workshop on Automatic Text Adaption (ATA); 2018. p. 1-11.

[11] Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Int Conf on New Methods in Language Processing; 1994. p. 44-9.

[12] Namer F. FLEMM : un analyseur flexionnel du français à base de règles. Traitement automatique des langues (TAL). 2000;41(2):523-47.

[13] Quinlan J. C4.5 Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.

[14] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

[15] Platt JC. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In: Advances in Kernel Methods - Support Vector Learning. MIT Press; 1998. .

[16] Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Int Conf on Machine Learning. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2001. p. 282-9. Available from: http://dl.acm.org/citation.cfm?id=645530.655813.

[17] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;7(8):1735-80.

[18] Rosenblatt F. The Perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review. 1958;65(6):386-408.