MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231045

TAXN: Translate Align Extract Normalize, a Multilingual Extraction Tool for Clinical Texts

Antoine NEURAZ[a,1], Ivan LERNER[a,b], Olivier BIROT[a], Camila ARIAS[a], Larry HAN[c], Clara Lea BONZEL[c], Tianxi CAI[c], Kim Tam HUYNH[a], and Adrien COULET[a]
[a]*Heka Team, Inria, INSERM Centre de recherche des Cordeliers, Université Paris Cité, Paris, France*
[b]*Department of Biomedical Informatics, Hôpital Européen Georges Pompidou, Hôpital Necker-Enfants Malades, APHP, Paris, France*
[c]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*

Abstract. Several studies have shown that about 80% of the medical information in an electronic health record is only available through unstructured data. Resources such as medical terminologies in languages other than English are limited and restrain the NLP tools. We propose here to leverage English based resources in other languages using a combination of translation, word alignment, entity extraction and term normalization (TAXN). We implement this extraction pipeline in an open-source library called “medkit”. We demonstrate the interest of this approach through a specific use-case: enriching a phenotypic dictionary for post-acute sequelae in COVID-19 (PASC). TAXN proved to be efficient to propose new synonyms of UMLS terms using a corpus of 70 articles in French with 356 terms enriched with at least one validated new synonym. This study was based on freely available deep-learning models.

Keywords. Natural language processing, named entity recognition, term normalization, implementation

1. Introduction

Several studies have shown that about 80% of the medical information in an electronic health record is only available through unstructured data [1]. To enable the secondary use of such information, extraction tools are required. Several tools exist already and provide variable performances [2-4]. These tools are often dedicated to English texts. However, there are also clinical texts written in other languages. Moreover, medical terminologies are more diverse and richer in English compared to other languages. For example, the Unified Medical Language System (UMLS) [5] contains only a few hundred thousand terms in French compared to a few millions in English. Therefore, entity linking towards the UMLS in English leads to a richer representation than in many other languages.

[1] Corresponding author: Antoine Neuraz, Antoine.neuraz@aphp.fr

Large resources are required to translate and maintain terminologies in different languages. In a heterogeneous and constantly evolving environment, reaching the goal of fully multilanguage terminologies seems out of reach. Instead of translating all terminologies, another method could be to translate medical texts into a language such as English, with rich terminologies, and perform the entity recognition on the translated texts. However, such approach leads to terms extracted in a language different than the original text, preventing a direct link between the text and the extracted entities.

In this work, we propose the translate, align, extract, normalize (TAXN) approach to easily extract entities on translated text while maintaining the link between the extracted entity and the original text. This is possible through a translation-alignment step that keeps track of the relations between original and translated words.

We demonstrate the interest of this approach through a concrete use-case: enriching a phenotypic dictionary for post-acute sequelae in COVID-19 (PASC).

All the tools required to perform this work is available through an open-source library called Medkit : https://github.com/TeamHeka/medkit.

2. Methods

Medkit is an open-source python library aiming a facilitating and accelerating the development and deployment of data processing pipelines. A specific attention has been paid to enable its usage in the context of healthcare through the detailed management of data provenance and non destructive processing. Data provenance is essential to ensure the traceability of data points created through processing pipelines. In the same spirit, non destructive processing allows to got back to the source of any data point and for example, display an entity extracted from text directly in the raw text.

Moreover, Medkit allows to build complex processing pipelines by chaining processing modules.

In this work, we built a three headed pipeline to extract medical entities from clinical texts (Figure 1).

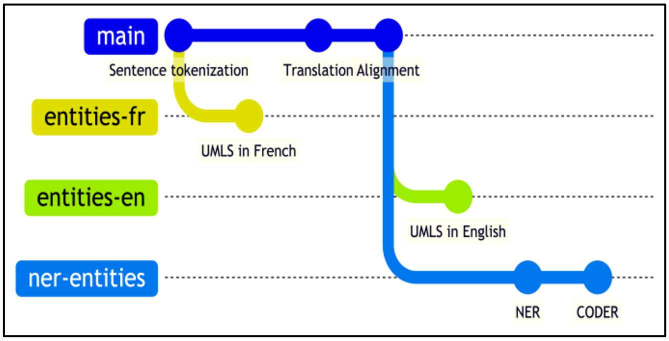

Figure 1. Description of the experiments.

We used 5 different processing modules for this work: 1) a rule based sentence tokenizer module; 2) a UMLS matcher, allowing a fuzzy matching on the UMLS metathesaurus, based on QuickUMLS [7]; 3) a deep-learning translation (Opus MT french-english [8]) and alignment (awesome Align [9]) module; 4) a named entity recognition (NER) module based on a pretrained Bert on N2c2 downloaded from the HuggingFace Hub (samrawal/bert-base-uncased_clinical-ner); 5) a UMLS normalizer module (CODER normalizer [9]) to normalize entities extracted via the NER module to UMLS terms.

The dataset used for this study came from a set of articles in French about long COVID. This corpus of 70 free articles was manually collected from the web.

Evaluation. We performed a comparison of the sets of CUIs identified through the different methods using UpsetR. A Manual evaluation of the correctness of the identified synonyms was also realized. For each potential synonyms, a manual review assessed the status, either correct or incorrect. We varied the number of minimal occurrences used to select a potential synonym between 1 and 50 and computed the corresponding accuracies.

3. Results

3.1. Sets of concepts with the different methods

In Figure 2, the comparison of sets of concepts extracted through the different methods is available. even if we can see some overlap between the different methods, it also appears that the 2 methods based on the translation-alignment models allow to obtain a majority of unique concepts (437 for the ner-entities and 417 for en-entities, respectively).

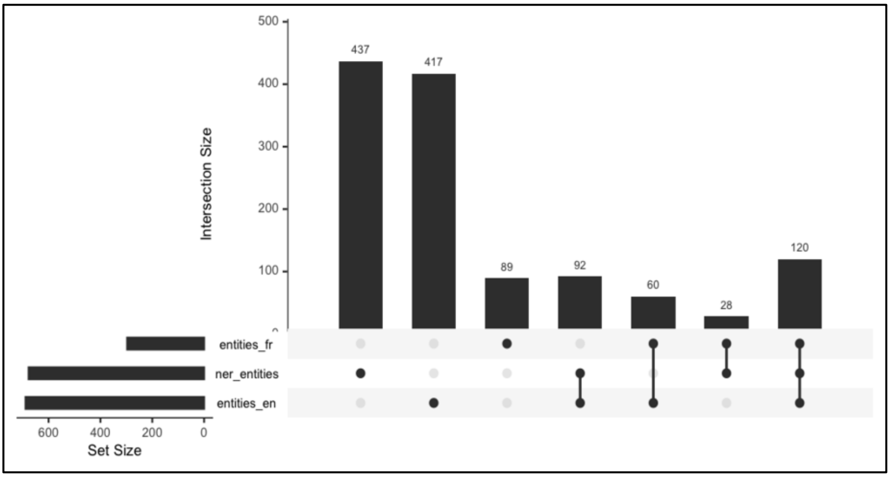

Figure 2. Comparison of the distribution of terms depending on their method of extraction. Only terms with at least 2 occurrences are displayed.

3.2. Quality of the synonym candidates

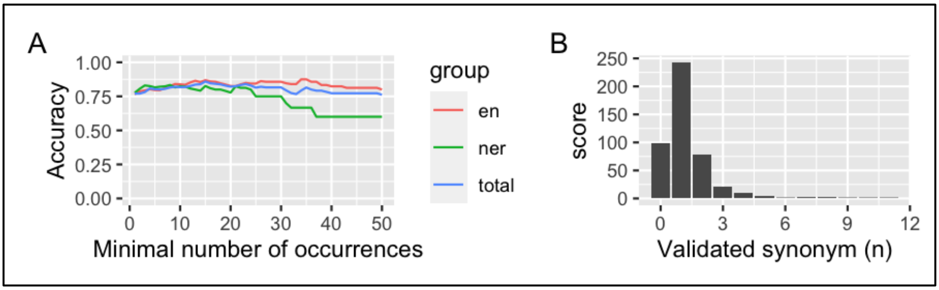

Figure 3. A/ Accuracy of the proposed synonyms depending on their number of occurrences. B/ number of validated synonyms (horizontally) and number of UMLS terms (vertically).

The accuracy of the proposed synonyms (manually evaluated) varied depending on the methodology but was overall at a median of 0.81 IQR[0.77, 0.82]. It was higher with the dictionary based approach entities-en with a median of 0.85 IQR[0.81, 0.86] compared witht the NER approach 0.75 IQR[0.60, 0.81]. Figure 3 Regarding the number of validated synonym per term: 98 terms had none, 243 had one and 113 had more than 1 (Figure 3). The manual review process took around 3 hours for the full set of candidate synonyms.

3.3. Qualitative examples

Table 1, provides examples of extracted synonyms with the different methods.

Table 1. Qualitative comparison of synonyms depending on the method.

entities_fr	entities_en	ner_entities
C0001617 Adrenal cortext hormones		
-	corticothérapie	-
corticoïdes		corticoïdes
corticostéroïdes		corticostéroïdes
C0002395 Alzheimer's Disease		
-	Alzheimer	Alzheimer
C0004358 Autoantibodies		
-	-	auto-anticorps
-	-	certains auto-anticorps
C0015624 Fanconi Syndrome		
-	dysfonction tubulaire proximale	dysfonction tubulaire proximale
syndrome de Fanconi	syndrome de Fanconi	-

4. Discussion

TAXN is an efficient method to extract potential synonyms from texts to enrich existing terminologies given that we could enrich 356 terms with at least one new synonym and 113 with more than one synonym in a small highly specialized corpus. The best performing approach in terms of accuracy was the dictionary based fuzzy matching. The NER approach allowed to extract more diverse synonyms as illustrated in the qualitative analysis. Overall, the median accuracy of the proposed synonyms was high enough to ensure a good usability. However, this approach still requires a human in the loop to validate the candidate synonyms reducing its potential for a fully automated use at large scale in this use-case. Another limit of this work resides in the number of deep learning models involved in the full pipeline using NER: currently 4 different transformer-based models are used. Such pipeline is therefore computationally expensive and will pose infrastructure problems for a production launch. Using more frugal models or combining different steps into a single model are improvement leads and future work opportunities. Interestingly, all the models used in this study are freely available and did not require a specific training.

5. Conclusions

TAXN proved to be efficient to propose new synonyms of UMLS terms using a corpus of 70 articles in French with 356 terms enriched with at least one validated new synonym. This study was based on freely available deep-learning models.

References

[1] Neuraz A, Lerner I, Digan W, Paris N, Tsopra R, Rogier A, Baudoin D, Cohen KB, Burgun A, Garcelon N, Rance B. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic. J Med Internet Res. 2020 Aug;22(8):e20773, doi: 10.2196/20773.

[2] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13, doi: 10.1136/jamia.2009.001560.

[3] Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. J Am Med Inform Assoc. 2017 Jul;24(4):841-4, doi: 10.1093/jamia/ocw177.

[4] Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, Box TL, DuVall SL, Patterson OV. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2022 Feb;2021:438-47.

[5] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Yearb Med Informatics. 1993 Jan;2(1):41-51, doi: 10.1055/s-0038-1637976.

[6] Dou ZY, Neubig G. Word alignment by fine-tuning embeddings on parallel corpora. arXiv. 2021 Jan, doi: 10.48550/arXiv.2101.08231.

[7] Soldaini L, Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, sigir 2016 Jul. p. 1-4.

[8] Tiedemann J, Thottingal S. OPUS-MT--Building open translation services for the World. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation 2020 Nov 1. European Association for Machine Translation. p. 479-80.

[9] Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform. 2022 Feb;126:103983, doi: 10.1016/j.jbi.2021.103983.