Taming large lexicons: translating clinical text using medical ontologies and sentence templates

François REMY
Internet and Data Science Lab
University of Ghent + imec
Gent, Belgium
francois.remy@ugent.be

Peter DE JAEGER
Innovation Center RADar
AZ Delta Hospitals
Roeselare, Belgium
peter.dejaeger@azdelta.be

Kris DEMUYNCK
Internet and Data Science Lab
University of Ghent + imec
Gent, Belgium
kris.demuynck@ugent.be

***Abstract*—One of the challenges of medical text translation is the usage of a large and specialized vocabulary, while the available parallel corpora are limited. In this presentation, we will show that including computer-generated parallel sentences deduced from a multi-lingual medical ontology in the training set of a Transformer-based translation model provides large benefits. Despite being trained exclusively on publicly available datasets, our model can achieve a performance level that is superior to the current state of the art in machine translation for the medical domain. We show this by comparing the precision and recall of MetaMap's medical concepts extraction [1] on a test set of clinical notes written in Dutch and subsequently translated to English, as well as by using more conventional quality metrics for translation models (BLEU, GPT-2 perplexity, cross-lingual cosine similarity).**

***Keywords*—translation, machine translation, clinical notes**

I. Introduction (*Heading 1*)

Several studies, including [2] and [3], have confirmed the importance of unstructured clinical notes in critical tasks for the healthcare industry, such as readmission or treatment outcome predictions. It is estimated that up to 80% of the relevant information in electronic health records (EHR) can only be found in the unstructured text typed by clinicians, despite the introduction of many structured-information sources over the years.

Two explanations for this are commonly cited: firstly, not every relevant piece of information can be mapped to an existing structured representation [4]; secondly, even when such representations exist, learning the existing coding systems and following through with their usage in practice is a source of friction that is sufficient to ensure that they will be routinely ignored by physicians, who find it easier to express themselves in natural languages [5].

Because of that, a wide variety of models and datasets have been developed to take advantage of these notes in different ways, with examples including entity recognition [1,6,7], natural language inference [8,9] and question answering [10].

Unfortunately, most resources usable for training remain available in English only. There are therefore few medical models of note in other languages. A solution to this problem would be to translate the resources currently available in English to other languages, such as Dutch, to be able to train models with similar capabilities in these languages. Another solution is to translate input from their native language to English. The latter is the approach evaluated in this paper.

So far, commercially available translation pipelines have not yet met the required quality standard to make EHR translation possible in a way that is compatible with the strict requirements of the medical domain. Patil and Devies [11] noted that "Google Translate has only 57.7% accuracy when used for medical phrase translations and should not be trusted for important medical communications", which is a surprisingly poor outcome given how popular the product is. While quality has improved in the recent years [12], there remains challenges with their usage in practice [13,14].

The main reason why these translation engines do not perform well for a domain is usually that there is insufficient train data for the domain [15]. Aligned clinical text is difficult to find [16]. Hiring human translator for medical text translation is usually listed at prices up to 10 times higher than for other types of text, making corpus creation costly [17].

Another reason is that the many medical texts come attached with strings, requiring anonymization and strict data retention policies [18]. These are not always compatible with the type of bulk processing large companies perform when building their translation pipelines.

This issue is compounded with the fact that medical text is highly specialized and has lots of ad-hoc terminology, and therefore requires many training examples to be accurately translated [19].

Finally, frequently-used translation datasets themselves contribute to quality issues due to their heavy reliance on automatically-aligned crawled data from the internet, which have been shown to be extremely noisy and of poor quality, with Caswell et al. [20] reporting that almost half of the translations found in such corpora are incorrect or otherwise unusable, on average.

In the remainder of this article, we will share how we tackled these various problems, as well as other practical issues we faced, to serve as a guide for developing high-quality translation pipelines for niche domains or languages. We then validate our results by looking at multiple quality metrics, including downstream tasks such as clinical term extraction.

II. Methodology

To achieve performance above the current state of the art in machine translation, we decided to focus on building a neural translation engine based on already-proven Marian MT [21], the technology stack behind Microsoft Bing Translator. This stack can be summed up as an Encoder-Decoder Transformer architecture with 6 layers on each side, static position embeddings, and a jointly-learnt vocabulary for the Dutch and English tokenization.

Our model was initialized using the OpusMT models [22] developed by the University of Helsinki. It was then finetuned using the HuggingFace library [23] on a V100 32Gb GPU.

For the development of this NMT model, various datasets were collected then assessed for suitability, semi-parallel corpora were mined for parallel examples using multilingual sentence embeddings models, data augmentation techniques were used to transform corpora comprising of unnatural phrases into more easily generalizable parallel sentences. We cover these operations in more detail in the following sections.

A. Finding parallel data for the biomedical domain

The first challenge we tackled was the collection of parallel corpora relating to medical data. Given the sizable cost of high-quality translations of such corpora, few of these exist; but a couple of sources can help us get started.

In the medical domain, drug notices are frequently legally-required to be translated in multiple languages for use in countries with multiple official languages, like Belgium. For many products sold in the EU, these translations are being registered at the European Union level, at the European Medicines Agency (EMA).

This is a good starting point, but these notices have a bias towards symptoms and diseases and do not cover medical procedures quite as much, which can be an issue. They are also aimed at a general public, and might try to avoid obscure terminology when alternatives are available, which clinical notes are not guaranteed to do. More datasets would be useful for achieving enough coverage.

Surprisingly, finding parallel corpora in the medical domain became easier when focusing on applications of NLP in the medical domain rather than on translation directly. An aligned corpus of Medline guidelines for clinicians was for example discovered while looking for multi-lingual named-entity recognition [24]. While not directly aimed at translation, these datasets can often align with that goal very well, and have the advantage of having been manually annotated, ensuring high quality.

Given the large breadth of the medical domain vocabulary, these two sets of corpora are not sufficient to achieve adequate coverage. Fortunately, the medical sector has developed over time many coding systems and ontologies, some of which have been partially or totally translated in multiple languages.

One such ontology, SnomedCT [25], was mined by us for parallel phrases. It is important to note that these ontologies are rarely aimed at translation, and while each concept can have multiple representations in any given language, these representations are not linked across languages. We therefore used Google's LaBSE model [26] to find best-matching pairs among the possibilities offered to us by the SnomedCT translation alternatives.

Given SnomedCT contains more than 700k concepts, using this mined corpus is an excellent way to improve vocabulary coverage. In addition, using this corpus biases the translation towards phrases that are easily recognizable by English NLP systems since they are part of the exact matches found in UMLS [27], a superset of SnomedCT. This lessens the burden on the English clinical term extraction model, by guiding the translation to well-known phrases.

Further investigation however revealed that adding these phrases straight to the training corpus did not yield the expected results while working with neural machine translation models, as they apparently had trouble translating known phrases when they were embedded inside sentences, even though they were well translated taken separately.

To overcome this and teach the model the role which these noun phrases can play in sentences (and thus boost generalization), a data augmentation technique was devised to generate several full sentences containing these terms, based on 50 manually crafted generic sentence templates.

For each concept in the ontology, parallel sentences were generated using 3 random templates out of these 50, by replacing a masked token by the concept names in Dutch and English respectively. The same LaBSE-matched pairs were used again, but including them in multiple sentences reinforced the translation model's ability to translate them correctly in context (see ablation studies).

B. Finding parallel data for spoken language

Another issue we noticed with clinical notes is that, unlike most written technical text, they feature many similarities with spoken text (ellipses, chaotic punctuation, fluid grammatical structure) while all the before-mentioned medical corpora are carefully redacted.

To address this challenge, our training set was expanded to include the OpenSubtitles dataset [28], which features a large variety of parallel subtitles (from movies to documentaries). Adding these subtitles improved the fluency of our translations visibly, although we had to take particular care in overcoming some biases stemming from this dataset (for example, most movies in the US use the Imperial units system, and the translation engine started translating the metric values into imperials, which is not a desired outcome).

C. Overcoming the lack of in-domain EHR parallel corpus

Despite the usage of the above-mentioned resources, it remains a fact that no in-domain training data (i.e. clinical notes) has been suggested so far. To overcome this, the frequently-used technique of back-translation was employed.

In particular, sentences of the MIMIC-III clinical notes corpus [29] and of the NIH Clinical Trials outcome dataset [30] were back-translated from English to Dutch using Google Translate, and added to the training dataset for a final finetuning (described in more detail in the next section). While these translations are not perfect, what matters is the target English.

This was motivated by our impression that Google Translate's main advantage over its competitors seemed not to be better understanding of the encoded Dutch, but rather a translation into more fluent English.

Adding in-domain sentences in English should thus improve the output quality, by guiding the language model of the decoder.

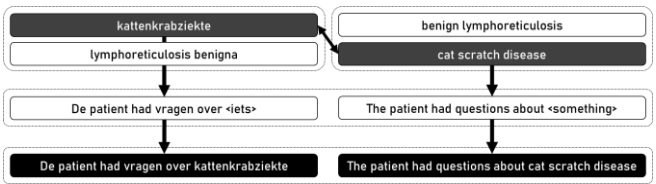

Fig. 1. Flowchart explaining the sentence-generation from Snomed CT

III. Experiments

To evaluate this model, the AZ Delta hospitals kindly provided us (post-training) with a small sample of anonymized text documents, extracted from their EHR. These texts are in the same medical domain as the train dataset, but feature more spoken-language phrasing, and the usage of punctuation in the text is not always consistent.

These short pieces of Dutch text are subsequently translated using the various systems being tested, and the quality of these translations is what is being reported in the results section below.

A sample of the pieces of text used is shown below *(dates have been redacted to increase the patient's privacy)*:

> **Datum 1:** Instabiele angor. Ostiale stenose LAD angularistak en stenose van meer dan 50% op de RCA. Stressischemie RCA en angularisgebied.
>
> **Datum 2:** Coronaire bypasschirurgie met lima naar angularistak en LAD en vene naar de ramus posterior descendens. Postoperatief eenmalig VKF.
>
> **Datum 3:** Blaasneoplasie waarvoor resectie.

Fig. 2. Example of test data used to evaluate the models

Note that this test set is different from the development set used for evaluating the model and choosing hyper-parameters values, which consisted of a left-out portion of the training set.

To evaluate the quality of the translations generated by the various systems, we decided to use the following four metrics, to avoid focusing on one particular aspect:

BLEU score on medical text translation: BLEU [31] is the most popular metric to judge translation quality. It compares the precision of a candidate translation in regard to a golden standard by counting their overlap in the n-gram domain. Terse and accurate translations often score well in BLEU score. Changes in word order at the global level do not affect the BLEU score too much, which enables for sentence-level word reordering at very low cost. Paraphrases or difference in language variations can however alter the BLEU score negatively. Despite these disadvantages, BLEU remains an important metric for translation rating, and translation models that perform well should also do well with this metric.

MetaMap concept extraction: MetaMap [1] is a frequently used system for medical concept extraction in text documents. It is based on the UMLS metathesaurus [27], and operates based on a set of rewriting rules followed by a precision/coverage scoring system for potential matches. Texts were first manually annotated by a medical ontology specialist, and the recall and precision of the MetaMap concept extraction are subsequently evaluated based on that golden standard. Because of the way MetaMap works, it is important for the generated translation to use phrases as close as possible to those usually found in medical text written in English to perform well on this test.

GPT-2 perplexity: GPT-2 [32] is an autoregressive language model trained on text extracted from the web. Its weights encode many bits of world knowledge, and is often used in transfer-learning scenarios, but in this case its usage as a language model is what we are interested in. Good translations should be plausible sentences in English, and therefore their perplexity in GPT-2 should be low. Translation models with a high fluency should score well with this metric, because they generate more plausible sentences than models that generate word-for-word translations.

XLM-R cosine similarity: XLM-R [33] is the state-of-the-art multi-lingual semantic sentence embedding system at the time of writing. An indication of how well sentences are translated would be the perceived similarity between their embeddings in the XLM-R space. Models that preserve the meaning of sentences should score well with this metric, but models that do not translate part of the sentences can also score high, so this metric should be used with caution.

IV. Results

All metrics confirm our model performs well for translating Dutch medical text into English. These results are summarized in Table 1.

TABLE I. Evaluation metrics of translation quality

Metrics / Model	Google	DeepL	Microsoft	Ours
BLEU (average)	64.3	61.8	58.5	69.1
MetaMap (recall)	82%	80%	69%	86%
GPT-2 (perplexity)	298	295	375	215

A. BLEU score on medical text translation

BLEU, the most used metric for translation, gives our model a 5 points advantage over Google on the test set. While BLEU is not meant to be averaged per-sentence, the relative ordering is what matters.

B. MetaMap concept extraction

Our most impressive result is an improvement of recall on a challenging medical concept extraction task based on MetaMap, where our model achieved a recall of 86% (against Google Translate's 82% and Microsoft Translator's 69%). We expect this to be in line with the downstream task performance obtained by our translations.

C. GPT-2 perplexity

When evaluated based on the GPT-2 perplexity of the translations, our model again performs the best on the test set (see Table 1).

The perplexity gap between our model and Google Translate (215 vs 298) seems to confirm that our model generates more fluent translation in the eyes of GPT-2.

That said, looking at the data manually seems to indicate that this only applies to specifically medical-domain constructs, as Google Translate seemed to remain more fluent overall.

D. XLM-R cosine similarity

When evaluated on the cross-language cosine similarity of the translations, all models seem to perform at about the same level, around 80% of similarity between sentence pairs.

The numbers even suggest it is possible that better translations yield a slightly worse average cross-lingual cosine similarity, but the difference (0.80 vs 0.81 vs 0.82 vs 0.80) is not significant given the size of our test set. We therefore chose not to include this metric in Table 1 given its lack of relevance, but still wanted to report our findings.

V. Ablation studies

To show the importance of the LaBSE matching of medical concepts found in SnomedCT, and the sentence-generation pipeline based upon these results, we ran two ablation studies.

NSS (No SnomedCT sentences): The first variant consists in a model trained on the same data as our best model, including the LaBSE-matched medical phrases from SnomedCT, but from which the computer-generated sentences including those matches were excluded. The results of this ablation show the gain in generalization obtained by generating sentences from SnomedCT terms.

NS (No SnomedCT): The second variant consists in a model trained on the same data as our best model, but from which both the SnomedCT matches found by LaBSE and their augmented sentences have been excluded. The results of this ablation study show the gain in vocabulary obtained by the addition of the SnomedCT matches.

As can be seen in the table below, disabling the sentence augmentation for SnomedCT concept pairs degrades the translation quality for MetaMap extraction by a factor of **3%**.

Excluding the SnomedCT pairs entirely from the train dataset further degrades the performance by an additional **4%**, at which point our model becomes worse than Google Translate for the task of concept extraction.

TABLE II. Evaluation metrics (ablation study)

Metrics / Model	Ours	NSS	NS
BLEU (average)	69.1	62.3	56.6
MetaMap (recall)	86%	83%	79%

These encouraging results demonstrate the value we were able to extract out of the SnomedCT translation dictionary we derived thanks to LaBSE. Based on these results, we decided to release this translation dictionary as well, because it can be useful in further experiments by other researchers.

VI. Conclusion

These encouraging results demonstrate the value we were able to extract out of the SnomedCT translation dictionary we derived thanks to LaBSE. Based on these results, we decided to release this translation dictionary as well, because it can be useful in further experiments by other researchers.

Thanks to our careful data collection and the generation of sentences based on the SnomedCT ontology, we were able to advance the state of the art for Dutch medical text translation, and to provide the first offline translation model achieving state-of-the-art results in the clinical domain.

Our model and the accompanying dictionary are now released on the HuggingFace repository. Taking advantage of the open licensing of the data used during training, both were released under a non-restrictive MIT License.

External resources

A. Translation Dictionaries

https://github.com/FremyCompany/snomed-translate-dictionaries

B. Translation Model

https://huggingface.co/FremyCompany/opus-mt-nl-en-healthcare

Acknowledgment

We would like to thank our partners AZ Delta for providing us with test examples to evaluate our model on, as well as the VLAIO organization for providing funding for the project from which this work is a part of.

REFERENCES

[1] Aronson, Alan R. 2001. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In AMIA Symposium, pages 17–21.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Monga, Kapila and Harpreet Singh. 2018. Unstructured data: An important piece of the healthcare puzzle. Journal Of AHIMA.

[3] Murphy, Kyle. 2013. Where is the value in physician notes, unstructured data? https://ehrintelligence.com/news/where-is-the-value-in-physician-notes-unstructured-data/

[4] Tewari, Anurag. 2014. Medical ontology: Big data big challenges. PMC, 02.

[5] Trapani, Marilyn. 2016. Why the ehr documentation burden needs to be solved (healthdata management). https://www.healthdata-management.com/articles/why-the-ehr-documentation-burden-needs-to-be-solved

[6] Neumann, Mark, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 319–327, Florence, Italy, August. Association for Computational Linguistics.

[7] Sung, Mujeen, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In ACL

[8] Romanov, Alexey and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1586–1596, Brussels, Belgium, October-November. Association for Computational Linguistics.

[9] Sarti, Gabriele. 2020. Biobert-nli (biobert model finetuned for natural language inference). https://huggingface.co/gsarti/biobert-nli.

[10] Ben Abacha, Asma, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 370–379, Florence, Italy, August. Association for Computational Linguistics.

[11] Patil, Sumant and Patrick Davies. 2014. Use of google translate in medical communication: evaluation of accuracy. BMJ, 349.

[12] Jackson, Jeffrey L, Akira Kuriyama, Andreea Anton, April Choi, Jean-Pascal Fournier, Anne-Kathrin Geier, Frederique Jacquerioz, Dmitry Kogan, Cecilia Scholcoff, and Rao Sun. 2019. The accuracy of google translate for abstracting data from non–english-language trials for systematic reviews. Annals of internal medicine, 171(9):677–679.

[13] Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. Information, Communication & Society, 24(11):1515–1532.

[14] Wetsman, Nicole. 2021. Google translate still isn't good enough for medical instructions. https://www.theverge.com/2021/3/9/22319225/google-translate-medical-instructions-unreliable

[15] Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. EAMT Proceedings, 04.

[16] Liu, Boxiang and Liang Huang. 2021. Paramed: a parallel corpus for english-chinese translation in the biomedical domain. BMC Medical Informatics and Decision Making, 21(1).

[17] Costowl. 2021. How much does a medical translation service cost? https://www.costowl.com/b2b/ translation-service-medical-cost.html.

[18] EMA. 2017. Data anonymisation - a key enabler for clinical data sharing. https://www.ema.europa.eu/ en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing en.pdf

[19] Jimenez-Crespo, Miguel Angel and Maribel Tercedor Sanchez. 2017. Lexical variation, register and explicitation in medical translation: A comparable corpus study of medical terminology in us websites translated into spanish. Translation and Interpreting Studies, 12(3):405–426.

[20] Caswell, Isaac, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoit Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets, 03.

[21] Junczys-Dowmunt, Marcin, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 129–135, Melbourne, Australia, July. Association for Computational Linguistics.

[22] Tiedemann, Jorg and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal.

[23] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October. Association for Computational Linguistics.

[24] Rebholz-Schuhmann, Dietrich, Simon Clematide, Fabio Rinaldi, Senay Kafkas, Erik M. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn, and Jan A. Kors. 2013. Entity recognition in parallel multi-lingual biomedical corpora: The clef-er laboratory overview. In Forner, Pamela, Henning Muller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pages 353–367, Berlin, Heidelberg. Springer Berlin Heidelberg.

[25] Schulz, Stefan and Gunnar O. Klein. 2008. Snomed ct – advances in concept mapping, retrieval, and ontological foundations. BMC Medical Informatics and Decision Making, 8(1).

[26] Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. https://arxiv.org/ abs/2007.01852.

[27] Bodenreider, Olivier. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(1):D267– D270, 01.

[28] Lison, Pierre and Jorg Tiedemann. 2018. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. https://www.opensubtitles.org/.

[29] Johnson, Alistair E.W., Tom J. Pollard, and Lu Shen. 2016. Mimic-iii, a freely accessible critical care database. In Scientific Data, volume 3. NIH.

[30] NIH. 2021. Clinicaltrials.gov - an online database of clinical research studies. https://clinicaltrials.gov.

[31] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

[32] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language models are unsupervised multitask learners.pdf

[33] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online, July. Association for Computational Linguistics