Evaluation of Machine Translation Methods applied to Medical Terminologies

Konstantinos Skianis
BLUAI
Athens, Greece
skianis.konstantinos@gmail.com

Yann Briand, Florent Desgrippes
Agence du Numérique en Santé
Paris, France
yann.briand@sante.gouv.fr
florent.desgrippes@gmail.com

Abstract

Medical terminologies resources and standards play vital roles in clinical data exchanges, enabling significantly the services' interoperability within healthcare national information networks. Health and medical science are constantly evolving causing requirements to advance the terminologies editions. In this paper, we present our evaluation work of the latest machine translation techniques addressing medical terminologies. Experiments have been conducted leveraging selected statistical and neural machine translation methods. The devised procedure is tested on a validated sample of ICD-11 and ICF terminologies from English to French with promising results.

1 Introduction

Medical terminologies are of essential importance for health institutions to store, organize and exchange all medical-related data generated in labs, hospitals and other healthcare entities. They are arranged systematically in dictionaries and lexicons, that follow specific structures and coding rules. In order to facilitate hierarchies and connections, the terms are represented by ontologies, enabling us to keep additional information (e.g. a family of diseases).

WHO International Classification of Diseases (ICD)[1] terminology is a diagnostic classification standard for epidemiology, clinical and research purposes. It is the most used medical dictionary across national health organizations worldwide. WHO is responsible to maintain the ICD editions for the English language. ICD-11 is the latest edition, adopted on May 25th, 2019. As the initial medical lexicons which contain these ontologies are created in English, there is an evident need for translation in other languages. This translation process can be expensive both in terms of time and resources, while the vocabulary and number of medical terms can reach high numbers and require health professional efforts for evaluation.

This work constitutes a generic, language-independent and open methodology for medical terminology translation. To illustrate our approach, which is based on automated machine translation methods, we will attempt to develop a first baseline translation from English to French for the ICD-11 classification. We also test on the International Classification of Functioning, Disability and Health (ICF) terminology[2].

First, we are going to investigate existing machine translation research studies concerning medical terms and documents, with a comparison of the relative methods. Next, we present our proposed methodology. Afterwards, we show our experiments and results. Last, we conclude with recommendations for future work.

2 Related Work

Translating medical terminologies has been a well-studied topic, with many approaches coming from machine translation. Traditional machine translation models first incorporated statistical models, whose parameters are set through the analysis of bilingual text corpora.

Statistical machine translation (SMT) Eck et al. (2004) investigated the usefulness of a large medical database (the Unified Medical Language System) for the translation of dialogues between doctors and patients using a statistical machine translation system. They showed that the extraction of a large dictionary and the usage of semantic type information to generalize the training data significantly improves the translation performance.

[1] https://icd.who.int/en

[2] http://bioportal.lirmm.fr/ontologies/ICF

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 59–69
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

	Resources	Type	Method	Languages
Nyström et al. (2006)	ICD-10, ICF, MeSH	SMT	Alignment	En-Swe
Deléger et al. (2010)	MeSH, SNMI, MedDRA 17, WHO-ART	SMT	Knowledge, Corpus	En-Fr
Laroche and Langlais (2010)	Wiki	SMT	Projection-based	Fr-En
Dušek et al. (2014)	EMEA, UMLS, MAREC	SMT	Domain	Multi
Silva et al. (2015)	SNOMED CT, DBPedia	Auto	Alignment	En-Por
Wołk and Marasek (2015)	EMEA	NMT	Encoder-Decoder	Pol-En
Arcan et al. (2016)	Organic.Lingua	SMT	Domain	En-(Ge, It, Sp)
Arcan and Buitelaar (2017)	ICD, Wiki	Both	Knowledge Base	En-Ge
Renato et al. (2018)	DeCS, Dicionario Medico, Wiki	SMT	Domain	Sp-Por
Khan et al. (2018)	UFAL, PatTR	NMT	Domain	En-Fr

Table 1: Summary of recent techniques for medical terms and texts translation.

Claveau and Zweigenbaum (2005) presented a method to automatically translate a large class of terms in the biomedical domain from one language to another; it is evaluated on translations between French and English. Their technique relies on a supervised machine-learning algorithm, called OSTIA (Oncina, 1991), that infers transducers from examples of bilingual term-pairs. Such transducers, when given a new term in English (respectively French), must propose the corresponding French (resp. English) term.

Later, Nyström et al. (2006) reports on a parallel collection of rubrics from the medical terminology systems ICD-10, ICF, MeSH, NCSP and KSH97-P and its use for semi-automatic creation of an English-Swedish dictionary of medical terminology. The methods presented are relevant for many other West European language pairs.

Deléger et al. (2009) presented a methodology aiming to ease this process by automatically acquiring new translations of medical terms based on word alignment in parallel text corpora, and test it on English and French. After collecting a parallel, English-French corpus, French translations of English terms were detected from three terminologies-MeSH, Snomed CT and the MedlinePlus Health Topics. A sample of the MeSH translations was submitted to expert review and a relatively high percentage of 61.5% were deemed desirable additions to the French MeSH. In conclusion, they successfully obtained good quality new translations, which underlines the suitability of using alignment in text corpora to help translating terminologies. Their method may be applied to different European languages and provides a methodological framework that may be used with different processing tools.

Neural machine translation (NMT) In recent years, NMT has emerged as the state-of-the-art approach. NMT uses a large artificial neural network which takes as an input a source sentence $(x_1, \ldots, x_m)$ and generates its translation $(y_1, \ldots, y_n)$, where x and y are source and target words respectively. Till recently, the dominant approach to NMT encodes the input sequence and subsequently generates a variable length translated sequence using recurrent neural networks (RNN) (Bahdanau et al., 2014; Sutskever et al., 2014). NMT differs entirely from phrase-based statistical approaches that use separately engineered subcomponents (Wołk and Marasek, 2015).

Domain adaptation In machine translation, domain adaptation can be applied when a large amount of out-of-domain data co-exists with a small amount of in-domain data.

Arcan and Buitelaar (2017) presented a performance comparison between SMT and NMT methods on translating highly domain-specific expressions, i.e. terminologies, documented in the ICD ontology from the medical domain. They showed that domain adaptation with only terminological expressions significantly improves the translation quality, which is specifically evident if an existing generic neural network is retrained with a limited vocabulary of the targeted domain. Last, they observed the benefit of subword models over word-based NMT models for terminology translation.

All previous work focus on training with specific terminologies. Although these methods are widely used, their vocabulary may be limited. Moreover, their size is not sufficient for training NMT methods, resulting in low translation performance.

To address these problems, Khan et al. (2018) trained NMT systems by applying transfer learning. Transfer learning falls under the umbrella of domain adaptation. In transfer learning the knowledge learned from a pre-trained existing model is

Terminology	Size	avg_len(en)	Incl	avg_len(en)
ICD-10	32474	5.49	7655	3.78
CHU Rouen HeTOP	202402	3.63	3892	3.69
ORDO	50425	6.2	3716	5.56
ACAD	47603	2.45	2394	1.84
MedDRA	23954	2.72	1739	2.33
ATC	5536	2.06	1588	1.11
MESH	29351	1.99	1460	1.69
ICD-O	3671	3.24	1122	2.88
DBPEDIA	912	1.78	381	1.85
ICPC	3046	7.09	235	2.26
ICF	3112	10.67	41	3.24
CLADIMED	4169	3.72	8	1.75
LOINC_2.66	91388	8.14	5	1.2
Total	499885	4.62	24242	3.35

Table 2: Reference terminologies and statistics regarding the validated sample of ICD-11. Number of sentences, average length in number of words (english corpus), number of included sentences in the validated sample of ICD-11, and their corresponding average length (number of words).

transferred to a new model. Specifically, the authors used an existing out-of-domain model trained on News data. Afterwards, they train their NMT system on the in-domain Biomedical'18 corpus[3].

Table 1 summarizes the related work on medical terms and texts translation, showing resources, family of machine translation approach, specific method used, languages studied and evaluation metrics, sorted by year.

3 Methodology

In the following section we describe the steps of our research methodology. First, a brief description of the terminologies and other corpora utilized is shown. Next, we describe the tools and libraries we have experimented with. Finally, the translation pipeline is presented.

3.1 Datasets

During our study we experimented upon numerous medical terminologies and datasets:

ATC (Anatomical Therapeutic Chemical, 2019). The ATC Classification System is a drug classification system that classifies the active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC), and was first published in 1976. Namely, the dataset includes descriptions on metabolism, blood, dermatological and other contents.

CLADIMED (CLADIMED, 2019) is a five levels classification for medical devices, based on the ATC classification approach (same families). Devices are classified according to their main use and validated indications. It was originally developed by AP-HP (hospitals of Paris).

ACAD (Académie de Médecine, 2019). The ”dictionnaire médical de l'académie de médecine” identifies terms used in health and defines them under the supervision of the French National Academy of Medicine.

ICD-O (World Health Organization, 2019). The International Classification of Diseases for Oncology (ICD-O) (1) has been used for nearly 35 years, principally in tumor or cancer registries, for coding the site (topography) and the histology (morphology) of the neoplasm, usually obtained from a pathology report.

MESH (Medical Subject Headings) (FR MESH, 2019) is a reference thesaurus in the biomedical field. The NLM (U.S. National Library of Medicine), which built it and updates it every year, uses it to index and query its databases, including MEDLINE/PubMed. INSERM, which has been the French partner of the NLM since 1969, translated the MeSH in 1986, and has been updating the French version every year since then. The bilingual version is often used as a translation tool, as well as for indexing and querying databases.

MedDRA (ICH, 2019) was developed in the late 1990s by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). It constitutes a rich and highly specific standardised medical terminology to facilitate sharing of regulatory information internationally for medical products.

ORDO (Vasant et al., 2014). The Orphanet Rare Disease Ontology (ORDO) is a structured vocabulary for rare diseases derived from the Orphanet database, capturing relationships between diseases, genes and other relevant features. Orphanet was established in France by the INSERM (French National Institute for Health and Medical Research) in 1997. ORDO provides integrated, re-usable data for computational analysis.

[3] https://www.statmt.org/wmt18/biomedical-translation-task.html

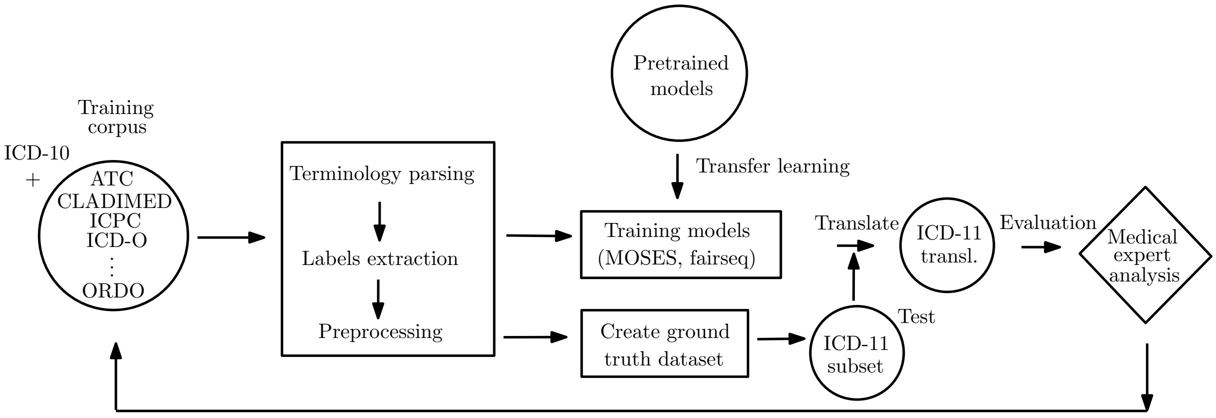

Figure 1: The proposed machine translation pipeline for ICD-11.

dbpedia (Auer et al., 2007). Through its API, dbpedia exposes multilingual fields and then can be used as a source to consolidate bi-lingual corpora.

ICD-10 (World Health Organization, 2016). ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. Work on ICD-10 began in 1983, endorsed by the Forty-third World Health Assembly in 1990, and it was first used by member states in 1994.

ICPC-2E (Verbeke et al., 2006). ICPC-2 classifies patient data and clinical activity in the domains of general/family medical practice and primary care, taking into account the frequency distribution of problems seen in these domains. It allows classification of the patient's reason for encounter, diagnostic, interventions, and the ordering of these data in an episode of care structure.

LOINC_2.66 (McDonald et al., 2003) is a widely used terminology standard for health measurements, observations, and documents.

CHU Rouen HeTOP is a large parallel corpus[4], including terminologies and ontologies in the domain of health, one of them being SNOMED CT[5].

ICF The International Classification of Functioning, Disability and Health (ICF), is a classification of health and health-related domains. ICF is the WHO framework for measuring health and disability at both individual and population levels.

In Table 2 we present the collection of medical terminologies and documents we explored during our research studies, as well as some statistics computed on them. We report size, average length of sentences in number of words, and number of sentences included in the validated sample of ICD-11.

3.2 Tools & libraries

Here we present publicly available tools that we used in our experiments. All the toolkits are written in Python, which offers a balance between complexity and usability. The Python community has increased dramatically during the past years, offering state-of-the-art methods in widely used libraries.

MOSES (Koehn et al., 2007) The MOSES tool software, is a phrasal-based probabilistic machine translation engine, which was used by many teams at the First Conference on Machine Translation (WMT16) (Bojar et al., 2016). Its base method includes word-alignment, phrase extraction and scoring during the training process.

fairseq (Ott et al., 2019) is a sequence modelling toolkit that allows researchers and developers to train custom models for translation, among other tasks. The toolkit offers a plethora of NMT models, like Long Short-Term Memory networks (LSTM) (Luong et al., 2015), Convolutional Neural Networks (CNN) (Dauphin et al., 2017; Gehring et al., 2017), as well as Transformer networks with self-attention (Vaswani et al., 2017; Ott et al., 2018).

Byte Pair Encoding (BPE) One of the most common problems in translating terminologies, including medical terminologies, are infrequent or unknown words, which the system has rarely or

[4] https://www.hetop.eu/hetop/
[5] http://www.snomed.org/

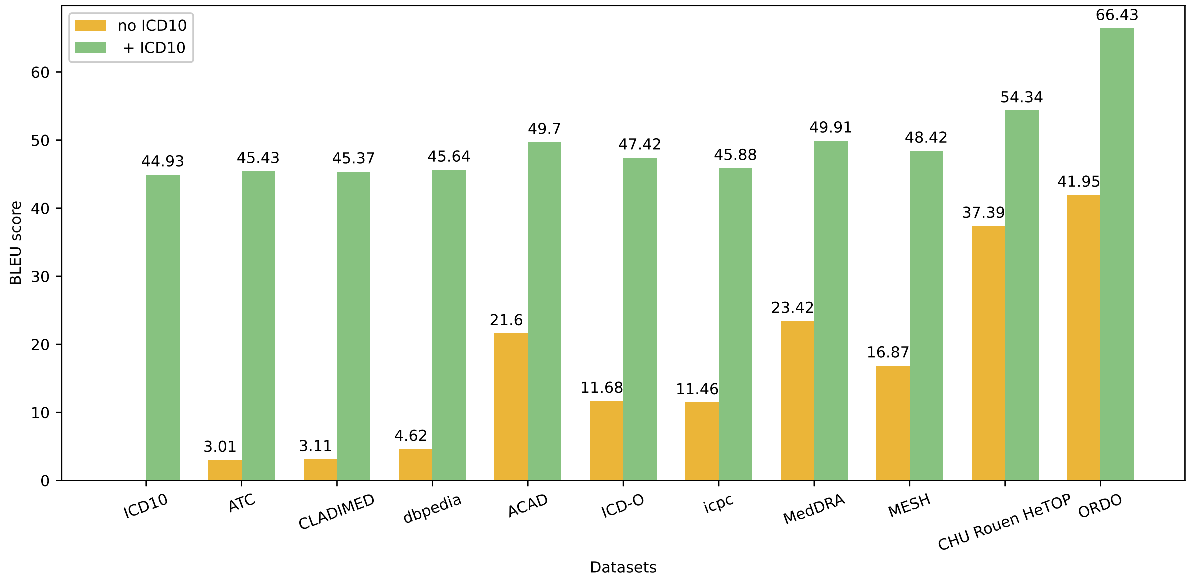

Figure 2: BLEU scores of MOSES on each dataset with and without ICD-10 in the training corpus using `multi-bleu.perl` by MOSES.

never seen. The effect is even more critical for NMT methods, where the vocabulary can not exceed the size of 50,000 or 100,000 words, due to the associated complexity. This limitation can be tackled by using subword units (BPE), a data compression technique (Sennrich et al., 2015). This step can be seen as part of preprocessing for the datasets, before training the models. We train our own BPE when no pre-trained model is used. In the transfer learning experiments, we use the provided BPE, as described in Section 4.

3.3 Dataset pipeline setup

An abstractive illustration of our proposed methodology is shown in Figure 1. Essentially, the pipeline can be split in five major parts: i) dataset & terminologies' search and retrieval, ii) parsing, extraction, preprocessing and extracting ground truth data, iii) model training, iv) translation and inspection, and v) evaluation and expert analysis.

Having access to the aforementioned datasets, we first applied terminology parsing. Next, we extracted the labels or descriptions, in order to form the corpus of parallel sentences. During the preprocessing step, we need to prepare the data for training the translation systems and perform tokenisation, truecasing and cleaning. For the NMT models, the BPE process is applied.

For ICD-11 given the fact that there is presently no human validated reference translation for French, we manually created one. The main objective of our work is to examine how fast and effective a translation to a newly created or updated medical terminology can be developed, to be given to medical experts for preliminary evaluation work.

Terminology	Size	avg_len(en)
ICD-11	123445	8.95
ICF	5920	10.79
Validated sample of ICD-11	24242	3.55

Table 3: Size (number of sentences) and average length in number of words (english corpus) for ICD-11, ICF and validated sample of ICD-11.

Our attempt offers the possibilities of speeding up the process of translating medical lexicons and documents, saving valuable human and computational resources. We evaluate our pipeline in two datasets: a sample of ICD-11 and the whole ICF terminologies. In the case of ICF terminology, we have access to both English and French medical experts validated versions. For ICD-11, since the French official version does not exist yet, we develop a method to evaluate and validate our results.

Through our studies, we discovered that a sample of the English ICD-11 terms can be found in existing French dictionaries. Thus, we can use these terms along with their French translation as already human-validated sentences. We end up having 24242 pairs in English and French that are already integrated in terminologies like ORDO, MESH_INSERM, LOINC_2.66 and others. Although, existing terms may as well require revi-

Method	Type	SacreBLEU ↑	BLEU ↑	METEOR ↑	TER ↓
MOSES no ICD10 (sys1)	SMT	39.92	35.61	33.84	50.61
MOSES only ICD10 (sys2)	SMT	45.84	39.16	35.18	45.22
MOSES dicts with ICD10 (sys3)	SMT	**65.59**	**57.50**	**46.20**	**28.62**
fairseq CNN no pre-trained (sys4)	NMT	51.02	42.93	38.85	38.98
fairseq CNN only pre-trained (sys5)	NMT	29.98	27.18	29.22	59.02
fairseq CNN finetuned on medical term/gies (sys6)	NMT	62.32	53.40	41.41	34.92
fairseq CNN finetuned on medical UFAL (sys7)	NMT	32.57	28.78	30.45	54.19

Table 4: SacreBLEU, BLEU, METEOR and TER scores on validated sample of ICD-11. Bold indicates best performance. SacreBLEU, BLEU and METEOR need to be maximized, while TER needs to be minimized.

Method	Type	SacreBLEU ↑	BLEU ↑	METEOR ↑	TER ↓
MOSES dicts with ICD10 (sys8)	SMT	50.40	42.82	38.74	38.50
fairseq finetuned on medical term/gies (sys9)	NMT	**60.82**	**52.46**	**42.97**	**32.59**

Table 5: Results on the ICD-11 24k sample, removed by the training dataset.

sion by a medical expert, the process indisputably accelerates the translation pipeline, compared to translating a terminology from scratch.

The automatic translation evaluation is based on the correspondence between the output and reference translation (ground truth/gold standard). We use popular metrics that cover several approaches:

- BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is calculated for individual translated segments (n-grams) by comparing them with a dataset of reference translations. Low BLEU score means high mismatch and higher score means a better match.

- SacreBLEU (Post, 2018) computes scores on detokenized outputs, using WMT (Conference on Machine Translation) tokenization and it produces the same values as the official script (`mteval-v13a.pl`) used by WMT.

- METEOR (Metric for Evaluation of Translation with Explicit ORdering) by Lavie and Agarwal (2007) includes exact word, stem and synonym matching while producing a good correlation with human judgement at the sentence or segment level (unlike BLEU which seeks correlation at the corpus level).

- TER (Translation Edit Rate) (Snover et al., 2006): the metric detects the number of edits (words deletion, addition and substitution) required to make a machine translation match exactly to the closest reference translation in fluency and semantics. High TER means high mismatch, while lower score means smaller distance from the reference text.

freq	sys3	sys6	len	sys3	sys6
1	0.8187	0.7858	-	-	-
2	0.8139	0.7626	<10	52.66	47.79
3	0.8263	0.7830	[10,20)	63.49	63.95
4	0.8429	0.7901	[20,30)	63.19	63.37
[5,10)	0.8521	0.8075	[30,40)	62.35	62.19
[10,100)	0.8714	0.8331	[40,50)	62.34	58.81
[100,1000)	0.7754	0.7749	[50,60)	59.64	59.82
≥1000	0.7773	0.7638	≥60	52.63	60.27

Table 6: Left: ICD-11 word accuracy analysis via `fmeasure` by frequency bucket. Right: sentence analysis by length bucket with BLEU metric for scoring.

Last, the translation is given to medical experts for analysis, recommending additional resources.

To the best of our knowledge, our work is one of the first that enables developing automatically a close to human-validated sample of a newly created or updated terminology. In Table 3 we present some statistics on our testing datasets.

4 Experiments & Results

In this section, we present the conducted experiments and obtained results. We selected two toolkits, due to their popularity and efficiency. MOSES represents the SMT tools, and fairseq represents the NMT domain. The summarized results of our experiments are visualized in Table 4. The traditional SMT model (sys3) manages to produce the best translation compared to the human validated sample, which consists mostly of short sentences. On the other hand, our best NMT model (sys6) performs slightly worse in total, but is better in longer sentences. The latter model (sys6) is finetuned on specialised medical terminologies, using as basis a largely pre-trained model on general do-

Method	Type	SacreBLEU ↑	BLEU ↑	METEOR ↑	TER ↓
MOSES dicts with ICD10 (sys3)	SMT	12.55	11.90	19.88	70.02
fairseq finetuned on medical term/gies (sys6)	NMT	**72.73**	**69.50**	**47.78**	**20.79**

Table 7: Results on translating the ICF terminology.

freq	sys3	sys6	len	sys3	sys6
1	0.3009	0.5323	-	-	-
2	0.2528	0.8251	<10	15.56	69.08
3	0.4284	0.8087	[10,20)	12.83	70.75
4	0.3315	0.8541	[20,30)	13.39	68.95
[5,10)	0.3501	0.8564	[30,40)	11.43	67.51
[10,100)	0.3812	0.8700	[40,50)	11.33	70.97
[100,1000)	0.5195	0.8761	[50,60)	6.44	69.93
≥1000	0.6644	0.8784	≥60	9.29	66.20

Table 8: Left: ICF word accuracy analysis via `fmeasure` by frequency bucket. Right: sentence analysis by length bucket with BLEU metric for scoring.

main corpora. In the next paragraphs we present our conducted experiments and results in detail.

MOSES We train our phrase-based translation system via MOSES, by building a 3-gram language model. First, we trained a model with all the medical terminologies excluding ICD-10 (sys1). We also experimented by using only ICD-10 (sys2) for training MOSES, reaching 44.93 in BLEU points. The model sys2 managed to perform better than any other dataset alone.

In order to identify the effectiveness of each terminology, we ran the translation process for each dataset separately, with and without ICD-10. Using only ATC, CLADIMED and dbpedia, resulted in poor performance, probably due to their specificity of included terms. Moreover, we observe that adding ICD-10 to all training datasets individually boosts the performance dramatically, as expected since many ICD-11 concepts come from ICD-10. Finally, training only on ORDO, we managed to reach a satisfying BLEU score. ORDO's effectiveness can be attributed to the large number of rare diseases it covers, which was one of the main improvements of ICD-11. The individual results are displayed in Figure 2.

Finally, we also trained an SMT model on the union of all the datasets. The model sys3 had the best performance, returning a high score of 65.59 SacreBLEU points, 57.50 BLEU points, 46.20 METEOR points and 28.62 TER points.

CNN trained on medical terminologies We trained a CNN model via fairseq on the medical terminologies we have gathered. The model (sys4) reports a very good performance with 51.02 SacreBLEU points and 42.93 BLEU points. Nevertheless, as the number of training epochs was relatively small (30 epochs), the model may present an even better performance if trained for more epochs.

fairseq's pre-trained CNN model fairseq provides online pre-trained models for many language pairs, offering multiple architectures, trained on large amount of textual data[6].

For our experiments we selected the freely available `wmt14.en-fr.fconv-py` model (Gehring et al., 2017). The convolutional neural network (CNN) was trained on the WMT'14 English-French dataset. The full training set consisted of 35.5M sentence pairs, where sentences longer than 175 words were removed. Last, a size of 40K BPE types was selected for the source and target vocabulary. We used the same BPE types for encoding the test datasets in both languages. The model required 8 GPUs for about 37 days for training, as stated in Gehring et al. (2017).

The fairseq pre-trained model reports a low BLEU score, with 27.18 points, due to its general out-of-domain training data. Moreover, fairseq fails to translate all sentences in a satisfying manner. The phenomenon of extraneous translations, like "HAUT DE LA PAGE" or "PEPUDU", can be confirmed by searching analogous patterns across the output. To address this issue, we finetuned fairseq's CNN on medical terminologies.

fairseq's CNN finetuned on medical terminologies The finetuned model (sys6) incorporates transfer learning as it continues training the pre-trained CNN model by fairseq (Gehring et al., 2017), described in the previous paragraph, on medical terminologies, presented in Section 3.1. The model (sys6) almost reached the performance of the SMT approach, with a performance of 62.32 SacreBLEU points and 53.40 BLEU points, while being close to sys3 in both METEOR and TER points as well. As we will also present later in our analysis paragraph, the finetuned model (sys6) is better in translating long sentences (len>50) than

[6] `https://github.com/pytorch/fairseq/tree/master/examples/translation`

Ground truth/Reference	**MOSES trained on medical terminologies (sys3)**	**fairseq CNN fine-tuned on medical terminologies (sys6)**
pied convexe congénital bilatéral	pied convexe congénital bilatéral (100)	astragale verticale congénitale bilatérale (0)
syphilis des ostia coronaires	syphilis des ostia coronaires (100)	maladie ostiale coronarienne syphilitique (0)
chute accidentelle de la personne portée	personne portée (9.56)	chute accidentelle de la personne portée (100)
maladie des inclusions microvilleuses	atrophie microvillositaire congénitale (0)	maladie des inclusions microvilleuses (100)

Table 9: Translation examples of our trained models on the verified sample of ICD-11, given by `compare-mt`. The number in parenthesis shows the sentence translation score in BLEU points compared to reference.

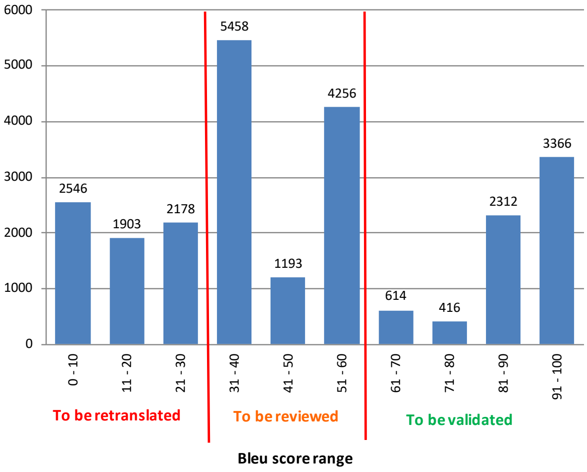

Figure 3: Sentence BLEU scoring on the 24k ICD-11 sample, categorized by a medical expert.

its MOSES rival (sys3), shown in Table 6. Our neural approach allows further training with no requirements.

fairseq's CNN finetuned on UFAL We also experimented on fine-tuning with the medical UFAL[7] dataset, a large medical domain corpus. The model (sys7) showed a performance of 28.78 BLEU points, being slightly better than using only the pre-trained CNN model. The low score can be attributed firstly to the short length nature of most ICD-11 sentences and secondly to the terminology syntax, which follows a specific structure. The medical UFAL consists mostly of long medical documents, which do not necessarily follow the typology of terminologies.

Removing the test sample from training As shown in Table 2, the validated sample of the ICD-11, which consists of 24k terms, is also included in the training dataset. Thus, we trained our two best architectures (sys3 & sys6) with removing the test set from the training corpus, creating two new models (sys8 & sys9). Table 5 presents their performance, showing that the neural model is far superior from the statistical approach.

Testing on ICF Since the validated sample of ICD-11 was mostly known sentences of short size belonging to terminologies, we believe that the SMT approach will perform worse than NMT in generalizing to unknown terms and sentences. To confirm this hypothesis, we tested on ICF, where the average length is 10.79 and thus larger than the ICD-11 average length. We tested our two best models, MOSES trained with all the datasets (sys3) and the finetuned CNN fairseq model (sys6) toward the ICF terminology. The finetuned CNN (sys6) performs far better than MOSES (sys3), by a large difference, with 69.50 BLEU points compared to a low 11.90 BLEU points, respectively. sys6 is also far superior to sys3 in terms of METEOR and TER points. The scores are presented in Table 7.

Analysis We also analyzed our best methods with `compare-mt`[8] (Neubig et al., 2019) to study their output. The tool offers aggregate scoring with BLEU and other metrics, word accuracy via `fmeasure`[9], sentence bucket and n-gram difference analysis. Our analysis is summarized in Table 6. We see that the MOSES model (sys3) performance ranges depending the frequency of terms, while our finetuned CNN (sys6) remains stable, regardless of the frequency. Looking at the right part of Table 6, sys3 performs worse when the length of terms increases significantly (len>50), but remains better than its rival (sys6) for length<10.

Regarding the ICF terminology, results are shown in Table 8. We clearly observe that the finetuned CNN (sys6) manages to translate well all ICF terms regardless of their frequency on words. Moreover, looking at the right part of Table 8, while sys6 provides promising results with both short and long terms, sys3 (the MOSES model) struggles to perform well, especially when the length increases.

[7] `https://ufal.mff.cuni.cz/ufal_medical_corpus`

[8] `https://github.com/neulab/compare-mt`

[9] `https://en.wikipedia.org/wiki/F1_score`

BLEU score range	English label	fairseq proposal (sys6) **Human translations**	Comments
0-0,2	Familial hypophosphataemic rickets	rickets hypophosphatémiques familiaux **Rachitisme familial hypophosphatémique**	Unknown word
	Adult-onset Still disease, buttock	apparition d’un adulte maladie mortelle, fessier **Maladie de Still survenant chez l’adulte, fesse**	Proper name misunderstood/not recognised
	common bile duct blunt injury	lésion de contour du canal biliaire commun **blessure contondante du canal cholédoque**	ambiguity of label (common)
0,21-0,5	Context of assault, gang rivalry	contexte de l’agression, rivalité entre gangs **Contexte d’agression, rivalité entre gangs**	Inappropriate insertion of article
	Barrett adenocarcinoma	adénocarcinome barrett **Adénocarcinome de Barrett**	proper name misunderstood missing coordination term
	Fracture of thumb bone	fracture du pouce **Fracture de l’os du pouce**	missing word
0,51-0,9	ureter cyst	cyste de l’uretère **kyste de l’uretère**	unknown word translated with editorialy very close term
	talipes equinovalgus	pied bot equinovalgus **talipes equinovalgus**	use of correct synonym
	Unintentional exposure to or harmful effects of oxazolidinediones	exposition non intentionnelle ou effets nocifs des oxazolidinediones **Effets nocifs ou exposition accidentelle à des oxazolidinediones**	word order and use of correct synonym

Table 10: Comparison of translation outputs with human translations.

We also present translation examples coming from our trained models, based on `compare-mt`. Table 9 shows four examples of the translation systems. The first two lines present a perfect translation coming from MOSES (sys3), while the last two lines show a perfect translation by the finetuned CNN model (sys6), due to transfer learning.

Next, we present a categorization of the translation BLEU scores on the 24k ICD-11 validated sample in Figure 3. The translations were studied by a medical expert, who extracted three categories using manually selected thresholds. A relatively small 27% of the translations required retranslation, a 45% needs to be reviewed and finally a 28% require to be just validated.

Last, a comparison translation outputs with human translations follows in Table 10. We present translation examples, given by the finetuned CNN model with medical terminologies (sys6), compare them with human translations, observing interesting linguistic phenomena. The comparison shows that as the BLEU score increases, the system outputs ”less acceptable” translations with cases like unknown words and ambiguities, to more ”acceptable” translations with cases like word order and correct synonym use.

5 Conclusion

In this work, an automated pipeline for translating and evaluating medical terminologies is presented. The pipeline is tested comparing different machine translation methods, to translate WHO ICD-11 and ICF terminologies from English to French. Over ten legacy medical terminologies along with ICD-10 are used for training the pipeline. A traditional MOSES SMT approach that manages to produce a good baseline translation is shown. We have tested NMT methods and found that finetuning largely pre-trained models like fairseq’s CNN on medical terminologies, incorporating transfer learning, can improve the quality of the translation. Last, we presented a simple method to generate automatically a test subset via existing terminologies.

The pipeline is adaptive to the typology of the studied terminology and it can be extrapolated easily to other languages for medical terminologies. The methodology enables researchers and healthcare end-users globally with a jump start approach that allows fast and effective translation of newly updated versions of terminologies.

For future work, using multilingual models (Liu et al., 2020) may omit the need for training multiple models in different languages. Last, additional medical datasets can be explored, not only for training but for creating larger validated corpora as well, following the constantly growing area of freely available language resources.

Acknowledgements

We would like to thank S. Darmoni from CHU Rouen / INSERM LIMICS (Paris) for sharing with us the HeTOP multilingual corpus that benefited the present study. We also thank the reviewers for their fruitful comments.

References

Académie de Médecine. 2019. Dictionnaire Médical de l'Académie de Médecine.

Anatomical Therapeutic Chemical. 2019. Atc.

Mihael Arcan and Paul Buitelaar. 2017. Translating domain-specific expressions in knowledge bases with neural machine translation. *arXiv preprint arXiv:1709.02184*.

Mihael Arcan, Mauro Dragoni, and Paul Buitelaar. 2016. Translating ontologies in real-world settings. In *International Semantic Web Conference*, pages 241–256. Springer.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 131–198.

CLADIMED. 2019. Classification des Dispositifs Médicaux (CLADIMED).

Vincent Claveau and Pierre Zweigenbaum. 2005. Translating biomedical terms by inferring transducers. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 236–240. Springer.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 933–941. JMLR.

Louise Deléger, Tayeb Merabti, Thierry Lecrocq, Michel Joubert, Pierre Zweigenbaum, and Stéfan Darmoni. 2010. A twofold strategy for translating a medical terminology into french. In *AMIA Annual Symposium Proceedings*, volume 2010, page 152. American Medical Informatics Association.

Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, and Daniel Zeman. 2014. Machine translation of medical texts in the khresmoi project. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221–228.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the unified medical language system. In *Proceedings of the 20th international conference on Computational Linguistics*, page 792. Association for Computational Linguistics.

FR MESH. 2019. Medical Subject Headings (MESH INSERM).

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

ICH. 2019. Medical Dictionary for Regular Activities.

Abdul Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. 2018. Hunter nmt system for wmt18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 655–661.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd international conference on computational linguistics*, pages 617–625. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *EMNLP*.

Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. 2003. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4):624–633.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

Mikael Nyström, Magnus Merkel, Lars Ahrenberg, Pierre Zweigenbaum, Håkan Petersson, and Hans Åhlfeldt. 2006. Creating a medical english-swedish dictionary using interactive word alignment. *BMC medical informatics and decision making*, 6(1):35.

Jose Oncina. 1991. *Aprendizaje de lenguajes regulares y transducciones subsecuenciales*. Ph.D. thesis, PhD thesis, Universidad Politécnica de Valencia, Valencia, Spain.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Kishore Papineni, 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alejandro Renato, José Castano, Maria Avila Williams, Hernan Berinsky, Maria Gambarte, Hee Park, David Pérez, Carlos Otero, and Daniel Luna. 2018. A machine translation approach for medical terms. pages 369–378.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Mario J Silva, Tiago Chaves, and Barbara Simoes. 2015. An ontology-based approach for snomed ct translation. In *ICBO*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Drashtti Vasant, Laetitia Chanas, James Malone, Marc Hanauer, Annie Olry, Simon Jupp, Peter N Robinson, Helen Parkinson, and Ana Rath. 2014. Ordo: An ontology connecting rare disease, epidemiology and genetic data.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Marc Verbeke, Diëgo Schrans, Sven Deroose, and Jan De Maeseneer. 2006. The international classification of primary care (icpc-2): an essential tool in the epr of the gp. *Studies in health technology and informatics*, 124:809.

Krzysztof Wołk and Krzysztof Marasek. 2015. Neural-based machine translation for medical text domain. based on european medicines agency leaflet texts. *Procedia Computer Science*, 64:2–9.

World Health Organization. 2016. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.

World Health Organization. 2019. WHO ICD Oncology (ICDO).