Investigating Massive Multilingual Pre-Trained Machine Translation Models for Clinical Domain via Transfer Learning

Lifeng Han[1], Gleb Erofeev[2], Irina Sorokina[2], Serge Gladkoff[2], and Goran Nenadic[1]
[1] The University of Manchester, UK
[2] Logrus Global, Translation & Localization
lifeng.han, g.nenadic@manchester.ac.uk
gleberof, irina.sorokina, serge.gladkoff@logrusglobal.com

Abstract

Massively multilingual pre-trained language models (MMPLMs) are developed in recent years demonstrating superpowers and the pre-knowledge they acquire for downstream tasks. This work investigates whether MMPLMs can be applied to clinical domain machine translation (MT) towards entirely unseen languages via transfer learning. We carry out an experimental investigation using Meta-AI's MMPLMs "wmt21-dense-24-wide-en-X and X-en (WMT21fb)" which were pre-trained on 7 language pairs and 14 translation directions including English to Czech, German, Hausa, Icelandic, Japanese, Russian, and Chinese, and the opposite direction. We fine-tune these MMPLMs towards English-Spanish language pair which did not exist at all in their original pre-trained corpora both implicitly and explicitly. We prepare carefully aligned clinical domain data for this fine-tuning, which is different from their original mixed domain knowledge. Our experimental result shows that the fine-tuning is very successful using just 250k well-aligned in-domain EN-ES segments for three sub-task translation testings: clinical cases, clinical terms, and ontology concepts. It achieves very close evaluation scores to another MMPLM NLLB from Meta-AI, which included Spanish as a high-resource setting in the pre-training. To the best of our knowledge, this is the first work on using MMPLMs towards clinical domain transfer-learning NMT successfully for totally unseen languages during pre-training.

1 Introduction

Multilingual neural machine translation (MNMT) has its root from the beginning of NMT era (Dong et al., 2015; Firat et al., 2016) but only made its first milestone when Google's end-to-end MNMT arrived (Johnson et al., 2017) where the artificial token was introduced for the first time for translation task at the beginning of the input source sentence to indicate the specified target language, e.g. "2en" as translating into English. This model used a shared word-piece vocabulary and enabled multilingual NMT through a single encoder-decoder model training. Google's MNMT also demonstrated the possibility of "zero-shot" translation as long as the languages to be translated from or to have been seen during the training stage, even though not explicitly. However, as the authors mentioned, Google's MNMT only allows translating between languages that have been seen individually as "source and target languages during some point, not for entirely new ones" in their many-to-many model, which was tested using the WMT14 and WMT15 data (Johnson et al., 2017). This set an obstacle to translating freshly new languages that do not exist in their pre-training stage. Then using the later developed NMT structure Transformer and BERT (Devlin et al., 2019; Vaswani et al., 2017), Facebook AI extended the coverage of multilingual translation into 50, 100, and 200+ languages via mBERT-50 (Tang et al., 2020), M2M-100 (Fan et al., 2021), and NLLB (NLLB Team et al., 2022) models. However, these models never address the issue of translating entirely new languages that do not exist in their pre-training stage, which sets an obstacle for MT applications in serving an even broader community.

In this work, we move one step forward towards domain-specific transfer-learning (Zoph et al., 2016) for NMT via fine-tuning an entirely new language pair that does not exist in the deployed multilingual pre-trained language models (MPLMs). The MPLMs we used are from Facebook AI (Meta-AI)' s submission to the WMT21 news translation

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 31–40
July 14, 2023 ©2023 Association for Computational Linguistics

task, i.e. "wmt21-dense-24-wide-en-X" and "wmt21-dense-24-wide-X-en" which were pre-trained for 7 languages Hausa (ha), Icelandic (is), Japanese (ja), Czech (cs), Russian (ru), Chinese (zh), German (de) to English (en), and backward (Tran et al., 2021). We use a well-prepared 250k pairs of English-Spanish (en-es) clinical domain corpus and demonstrate that not only it is possible to achieve successful transfer-learning on this explicit new language pair, i.e. the Spanish language is totally unseen among the languages in the MPLM, but also the domain knowledge transfer from general and mixed domain to the clinical domain is very successful. In comparison to the massively MPLM (MMPLM) NLLB which covers Spanish as a high-resource language at its pre-training stage, our transfer-learning model achieves very close evaluation scores in most sub-tasks (clinical cases and clinical terms translation) and even wins NLLB in ontology concept translation task by the metric COMET (Rei et al., 2020) using ClinSpEn2022 testing data at WMT22. This is a follow-up work reporting further findings based on our previous shared task participation (Han et al., 2022a) and pre-print (Han et al., 2022b).

2 Related Work

Regarding the early usage of special tokens in NMT, Sennrich et al. (2016) designed the token T from Latin Tu and V from Latin Vos for familiar and polite indicators attached to the source sentences towards English-to-German NMT. Yamagishi et al. (2016) designed tokens <all-active>, <all-passive>, <reference> and <predict> to control of voice of Japanese-to-English NMT; either they are active, passive, reference aware or prediction guided. Subsequently, Google's MNMT system designed target language indicators, e.g. <2en> and <2jp> controlling the translation towards English and Japanese respectively (Johnson et al., 2017). Google's MNMT also designed mixed target language translation control, e.g. $(1\text{-}\alpha)\langle 2\text{ko}\rangle + \alpha\langle 2\text{jp}\rangle$ tells a mixed language translation into Korean and Japanese with a weighting mechanism. We take one step further to use an existing language controller token from a MPLM as a pseudo code to fine-tune an external language translation model, which was entirely not seen during the pre-training stage.

Regarding transfer-learning applications for downstream NLP tasks other than MT, Muller et al. (2021) applied transfer learning from MPLMs towards unseen languages of different typologies on dependency parsing (DEP), named entity recognition (NER), and part-of-speech (POS) tagging. Ahuja et al. (2022) carried out zero-shot transfer learning for natural language inference (NLI) tasks such as question answering.

In this paper, we ask this research question (RQ): Can Massive Multilingual Pre-Trained Language Models Create a Knowledge Space Transferring to Entirely New Language (Pairs) and New (clinical) Domains for Machine Translation Task via Fine-Tuning?

3 Model Settings

To investigate into our RQ, we take Meta-AI's MNMT submission to WMT21 shared task on news translation, i.e. the MMPLM "wmt21-dense-24-wide-en-X" and "wmt21-dense-24-wide-X-en" as our test-base, and we name them as WMT21fb models (Tran et al., 2021)[1]. They are conditional generation models from the same structure of massive M2M-100 (Fan et al., 2021) having a total number of 4.7 billion parameters which demand high computational cost for fine-tuning. WMT21fb models were trained on mixed domain data using "all available resources" they had, for instances, from historical WMT challenges, large-scale data mining, and their in-domain back-translation. Then these models were fine-tuned in news domain for 7 languages including Hausa, Icelandic, Japanese, Czech, Russian, Chinese, German from and to English.

The challenging language we choose is Spanish, which did not appear in the training stage of WMT21fb models. The fine-tuning corpus we use is extracted from MeSpEn (Villegas et al., 2018) clinical domain data, of which we managed to extract 250k pairs of English-Spanish segments after data cleaning. They are from IBECS-descriptions, IBECS-titles, MedlinePlus-health_topics-titles, MedlinePlus-health_topics-descriptions,

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/wmt21

Рис. 1: (Figure:) Difference of Google's Multi-lingual NMT Bridge Models (left) and Our Transfer-Learning Model (right).

Pubmed-descriptions, Scielo-descriptions, and Scielo-titles.

To implement the fine-tuning, we use the <2en> token for translating from Spanish to English, and <2ru> (originally to Russian) pseudo token for translating towards English-to-Spanish (en2es) [2]. The difference between our transfer-learning NMT model and Google's MNMT can be shown in Figure 1, right vs left. In Google's MNMT model, it can only translate "new language pairs" that are not explicitly seen but implicitly seen, e.g. bridging language pairs (Ukrainian-to-English and English-to-Russian ⇒ Ukrainian-to-Russian), or language pairs that have been seen as source (Korean) and target (Portuguese) somewhere. In our transfer-learned NMT, Spanish was not among the trained languages at all.

In comparison, we deploy another MMPLM from Meta-AI, i.e. the "No-Language-Left-Behind (NLLB)" which was trained on 204 languages including Spanish as one of their high-resource ones (NLLB Team et al., 2022). NLLB full model is a massive size Transformer having 55 billion parameters and we use its distilled version NLLB-200-distilled [3], which still has 1.3 billion parameters. Fine-tuning is carried out on NLLB using the same 250K ES-EN corpus.

3.1 Model Parameters in Detail

Some fine-tuning parameters for NLLB-200-distilled (NLLB Team et al., 2022) are listed below:

- batch size = 24
- gradient accumulation steps = 8
- weight decay = 0.01
- learning rate = 2e-5
- number of training epochs = 1
- encoder-decoder layers = 24+24
- Activation function (encoder/decoder) = ReLU

The Parameters for fine-tuning WMT21fb model are the same as for the NLLB-200, except for the batch size which is set as 2, which is because the model is too large and we got an OOM error if the batch size is set above 2. More details on M2M-100 for Conditional Generation structure (Fan et al., 2021) we used can be find in Figure 2.

4 Model Evaluations

4.1 Testing Corpus from Clinical Domain

We used the official testing corpus from ClinSpEn2022 shared task affiliated to Biomedical-MT at WMT22. ClinSpEn2022 aims at developing clinical domain machine translation on Spanish-English language pair[4], which is hosted in CodaLab (Pavao et al., 2022) [5].

[2] using <2es> token will result into errors since Spanish was actually not used in the WMT21fb PLMs

[3] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[4] https://temu.bsc.es/clinspen/

[5] https://codalab.lisn.upsaclay.fr/competitions/6696

```
M2M100ForConditionalGeneration(
  (model): M2M100Model(
    (shared): Embedding(128009, 2048, padding_idx=1)
    (encoder): M2M100Encoder(
      (embed_tokens): Embedding(128009, 2048, padding_idx=1)
      (embed_positions): M2M100SinusoidalPositionalEmbedding()
      (layers): ModuleList(
        (0): M2M100EncoderLayer(
          (self_attn): M2M100Attention(
            (k_proj): Linear(in_features=2048, out_features=2048, bias=True)
            (v_proj): Linear(in_features=2048, out_features=2048, bias=True)
            (q_proj): Linear(in_features=2048, out_features=2048, bias=True)
            (out_proj): Linear(in_features=2048, out_features=2048, bias=True)
          )
          (self_attn_layer_norm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
          (activation_fn): ReLU()
          (fc1): Linear(in_features=2048, out_features=16384, bias=True)
          (fc2): Linear(in_features=16384, out_features=2048, bias=True)
          (final_layer_norm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
        )
        (1): M2M100EncoderLayer(
```

Рис. 2: (Figure:) M2M-100 Model Structure For Conditional Generation Encoder: Samples and Parameters

Task-I: Clinical Cases (CC) EN→ES						
MT fine-tuning	in.es?	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Clnical-NLLB	Yes	37.74	0.6273	0.4081	0.3601	0.6193
Clinical-WMT21fb	No	34.30	0.5868	0.3448	0.3266	0.5927
Task-II: Clinical Terms (CT) EN←ES						
MT fine-tuning	in.es?	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Clinical-NLLB	Yes	28.57	0.5873	1.0290	0.2844	0.6710
Clinical-WMT21fb	No	24.39	0.5840	0.8584	0.2431	0.6699
Task-III: Ontology Concept (OC) EN→ES						
MT fine-tuning	in.es?	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Clinical-NLLB	Yes	41.63	0.6072	0.9180	0.3932	0.7477
Clinical-WMT21fb	No	40.71	0.5686	0.9908	0.3859	0.7199

Таблица 1: (Table:) Evaluation Scores using Five Official Metrics from ClinSpEn2022 Benchmark on Two Models. The column "in.es" means if the original pre-trained model included the Spanish language before fine-tuning/transfer-learning.

There are three sub-tasks: 1) Clinical Cases (CC): on 202 COVID-19 clinical case reports; 2) Clinical Terms (CT): using more than 19K parallel terms extracted from biomedical literature and electric health records (EHRs); 3) Ontology Concepts (OC): using more than 2K parallel concepts from biomedical ontology. The translation direction on these three sub-tasks are EN→ES, EN←ES, and EN→ES respectively.

4.2 Evaluation Metrics

The official evaluation metrics used by ClinSpEn2022 shared task are METEOR (Banerjee and Lavie, 2005), SacreBLEU (Post, 2018), COMET (Rei et al., 2020), BLEU-HF (HuggingFace) (Papineni et al., 2002), and ROUGE-L-F1 (Lin, 2004). Among these, METEOR is a metric using both precision and recall not only on word surface level but also introducing paraphrasing features. COMET was proposed recently by taking advantage of cross-lingual PLMs using knowledge from both source and target languages. ROUGE was originally designed for text summarisation evaluation using n-gram co-occurrences, while ROUGE-L added the Longest Common Subsequence (LCS) feature from translation study.

The reporting of BLEU metric scores has certain uncertainty, which is caused by some parameter settings when using BLEU metric including number of references, length penalty computation on multi-references, maximum n-gram, and smoothing applied to 0-count n-grams. To address these issues, SacreBLEU added some constrains while using BLEU metric. These include the applying of its own metric-internal pre-processing for detokenised system outputs, the avoiding of user handling reference set via automatically downloading from WMT, and the export of a summary on settings used.

4.3 Evaluation Scores

We present the MT evaluation scores using five official metrics from ClinSpEn2022 shared task on the three sub-tasks in Table 1, for translating clinical cases, clinical terms, and clinical concepts. The two fine-tuned models are clinic-NLLB which is achieved by domain fine-tuning and clinic-WMT21fb which is a domain fine-tuning plus transfer-learning model to a new language space.

On Task 1 and 2, Clinical-WMT21fb has very comparable evaluation scores to clinical-NLLB, even though it only used 250k pairs es-en sentences for fine tuning without seeing any en-es or Spanish language at all during pre-training. In contrast, clinical-NLLB used a large amount of Spanish data for its pre-training phase. On Task 3, the evaluation scores of these two models are even closer on BLEU and SacreBLEU, especially the clinical-WMT21fb wining COMET metric over clinical-NLLB (0.9908 vs 0.9180).

This experimental result shows that with a carefully prepared certain amount of fine-tuning data, e.g. 250k pair of sentences, the MMPLMs are capable to create a semantic knowledge space transferring to an entirely new (external) language pair for NMT task in a new domain, i.e. clinical domain. This answers our RQ set up in the beginning of this investigation.

4.4 Human Evaluation

We looked into three sub-task translation outputs from the model clinical-WMT21fb. It shows that for the EN←ES translation task, i.e. the sub-task 2 clinical term translation, the output file is totally file with only English tokens. On the other two sub-tasks, i.e. the clinical cases and ontology concept translation, which have the translation direction EN→ES, there are some Russian tokens in the output, not only Spanish tokens. However, the Russian tokens in the Spanish sentences are not nonsense, instead proper translations of entities and words. The entire test set of these two sub-tasks is very large around 300K sentences/segments, and there are only 12K lines of them (4%) have Russian tokens. So we have fine-tuned the model in EN-RU direction on EN-SP data, and it translates well into Spanish! But if there isn't a suitable Spanish token in the generation model, it takes a Russian token.

We also looked into the translation outputs from clinic-NLLB model for error analysis using two native Spanish speakers, one of them having a PhD degree in biomedical NLP field and the other having a Master degree in translation studies. The error analysis shows that some of the translation errors come from very literal translation, and others come from gender related mistakes. This suggests that the massively pre-trained MLM is still not there to capture the differences of linguistic features among pre-trained languages.

5 Discussion

5.1 On Automatic Metrics

We had more thoughts on the automatic evaluation settings and outputs, especially on the COMET metric in comparison to others.

Firstly, the closeness of most automatic metric scores does not necessarily mean that the translation outputs are very good. Most metrics only measure the linguistic proximity of outputs to the "gold standard of reference".

Secondly, COMET is a reference-less metric taking advantage of cross-lingual PLMs using knowledge from both source and target languages. This has pros and cons: a) it might be able to capture the semantic relatedness without seeing the same language tokens, even in the same sequence/sentence; b) also due to this, it is not able to distinguish foreign language tokens in the translation output, which normally shall receive a penalty in evaluation scores. This also inspires another research topic, i.e. shall we really punish the foreign or mixed-language tokens in the translation output in all evaluation conditions, or it shall depend on the situation of the output applications? This has an echo to Google's zero-shot MNMT model (Johnson et al., 2017) when the mixed language tokens are used for translation model, e.g. (1-α)<2KO>+α<2JP> resulting in mixed tokens of Korean and Japanese in the output translation but they are semantically correct tokens.

In a situation when users want only the Spanish translation output, 4% of Russian tokens in the Spanish translation should surely receive a penalty in the quality evaluation setting. The COMET metric will fail this mission, and professional human evaluation is always much needed for trustworthiness. However, in a situation to measure the models' cross-lingual capability on semantic preservation for direct output, or as input into other ML models, is it better to generate NULL or meaningless tokens or random translations in the target language, or to choose semantically correct foreign tokens when the model does not know how to predict the exact correct target tokens? This inspires us to think again about the evaluation setting on different tasks.

5.2 MT System Output Examples

We present the MT system output examples from both clinical-WMT21fb and clinical-NLLB-200 for three tasks in Figure 3, 4, and 5. In these figures, the green colour is for the "preferred translations" while the orange colour is for "both sounds good". The annotations were firstly marked by one of the two human evaluators we have, and then verified by the second native Spanish speaker.

From these sampled MT outputs, the model clinical-WMT21fb sometimes outperforms clinical-NLLB-200, and vice versa. For instance, in the concept translation (Figure 5), the English concept "Abnormality of body height" (ont_1) is better translated by transfer-learned model into "Anomalía de la altura corporal" than "Anomalías de la talla corporal" by clinical-NLLB, since "altura" means "height" while "talla" actually means "size" which is not accurate. We will carry out a systematic human evaluation in a larger sample size.

Regarding rare Russian tokens from the language-transferred model, in Task-1, "Вскоре" from clinical-WMT21fb in line_n 4 means "soon", even though it is a Russian token, i.e. non-Spanish token. In Task-3, "Тип" in "Тип autosómico dominante" means "type of" from ont_11 which is a meaningful Russian token.

6 Conclusion and Future Work

We investigated if real transfer-learning NMT is possible using massive multilingual pre-trained LMs (MMPLMs) to translate external languages that are unseen at all in the training phase. We used Meta-AI's mixed domain multilingual PLMs (WMT21fb) as our test base, 250K well-prepared EN-ES clinical data as fine-tuning corpus, and <2ru> as pseudo-code for new language (out-of-en) fine-tuning. We tested the fine-tuned model on ClinSpEn2022 clinical domain shared task data, and the results show that this fine-tuning is successful, which achieves very comparable scores to Meta-AI's MMPLM NLLB model, which had Spanish in the training phase as a high-resource setting. We think this demonstrates that the Hyper-Transformer model from WMT21fb does build a language-independent "semantic space" that allows one to understand a different language and correctly construct a totally different language model when fine-tuned on the language which was absent and different from the languages it was trained upon. This finding can be very useful for future clinical knowledge transformation, e.g. from existing high-resource languages to low-resource languages, such that clinicians from low-resource language speakers can also benefit from AI-supported decision-making. The well-trained clinical models based on properly translated resources can also potentially support patients' self-diagnoses and self-care in originally scarce resource settings.

There are many future works that can be carried out based on the findings from this work. Firstly, we plan to carry out an extensive human-expert-based evaluation, e.g. using HOPE metric (Gladkoff and Han, 2022), looking into the differences between

doc_n	line_n	**Transfer-learning**: clinical-WMT21fb:en2es
doc_15976	0	Hombre de 58 años de edad, de raza caucásica, con diagnóstico de EP predominante en temblor a los 44 años de edad.
doc_15976	1	Agonistas dopaminérgicos y tratamiento con levodopa permitieron un buen control sintomático.
doc_15976	2	A los 48 años de edad fue diagnosticado VIH en una prueba **rutinaria**.
doc_15976	3	Seis años después, aunque permaneció asintomático, el recuento de CD4 alcanzó 209 células/µl y se inició TARGA.
doc_15976	4	Вскоре, después, aparecieron síntomas gastrointestinales severos (náuseas, vómitos y diarrea) y discinesias a dosis **pico**, que se atribuyeron a las interacciones farmacocinéticas entre levodopa y TARGA.
doc_15976	5	Inicialmente, la levodopa se redujo a costa de un control subóptimo de la **EP**, pero posteriormente el **tratamiento antirretroviral** ha de suspenderse debido a discinesias intolerables.
doc_15976	6	Tras 3 años de buen control sintomático de la EP y infección por VIH asintomática, el paciente comenzó a sufrir fuertes fluctuaciones motrices con distonía de mañana y discinesias de dosis máxima.
doc_15976	7	En el momento de considerarse STN-DBS, estaba en tratamiento con levodopa de liberación inmediata y controlada y ropinirol, totalizando una dosis diaria de levodopa equivalente de 1.250 mg.
doc_15976	8	PD estuvo en estadio 3 de Hoehn-Yahr durante la medicación, y la puntuación UPDRS-III fue 78 sin medicación y 18 tras la ingesta de levodopa.

doc_n	line_n	**Fine-tuning**: clinical-NLLB:en2es
doc_15976	0	Un hombre de 58 años de edad, de raza caucásica, fue diagnosticado de EP predominante en temblor a los 44 años.
doc_15976	1	Los agonistas de dopamina y el tratamiento con levodopa permitieron un buen control sintomático.
doc_15976	2	A los 48 años, fue diagnosticado **de VIH** en una prueba de rutina.
doc_15976	3	Seis años después, aunque **permanecía** asintomático, el recuento de CD4 había alcanzado 209 células/µl, y se inició la TARGA.
doc_15976	4	Poco después, se presentaron síntomas gastrointestinales graves (náuseas, vómitos y diarrea) y discinesias de dosis máxima, atribuidas a interacciones farmacocinéticas entre levodopa y TARPA.
doc_15976	5	Inicialmente, la levodopa se redujo a costa de un control subóptimo de la **EP**, pero posteriormente se tuvo que suspender la TARHA por las discinesias intolerables.
doc_15976	6	Tras 3 años de buen control sintomático de la EP y infección asintomática por el VIH, la paciente comenzó a sufrir de fluctuaciones motoras **severas** con distonía **matinal** y discinesias de dosis máxima.
doc_15976	7	Para el momento de la consideración de STN-DBS, estaba en levodopa y ropinirol de liberación inmediata y controlada, con una dosis equivalente diaria de 1.250 mg.
doc_15976	8	La EP se encontraba en estadio 3 de Hoehn-Yahr mientras estaba en tratamiento, y la puntuación UPDRS-III fue de 78 fuera de tratamiento y de 18 tras el consumo de levodopa.

doc_n	line_n	text:src:English
doc_15976	0	A 58-year-old Caucasian man was diagnosed with tremor-predominant PD at the age of 44 years.
doc_15976	1	Dopamine agonists and levodopa therapy allowed a good symptomatic control.
doc_15976	2	By the age of 48 years, he was diagnosed with HIV on a routine testing.
doc_15976	3	Six years later, although he remained asymptomatic, the CD4 count had reached 209 cells/µl, and HAART was started.
doc_15976	4	Soon after, severe gastrointestinal symptoms (nausea, vomiting, and diarrhea) and peak-dose dyskinesias emerged, which were attributed to pharmacokinetic interactions between levodopa and HAART.
doc_15976	5	Initially, levodopa was reduced at the cost of suboptimal control of PD, but afterwards HAART had to be discontinued because of intolerable dyskinesias.
doc_15976	6	After 3 years of good symptomatic PD control and asymptomatic HIV infection, the patient began to suffer from severe motor fluctuations with morning off dystonia and peak-dose dyskinesias.
doc_15976	7	By the time STN-DBS was considered, he was on immediate and controlled-release levodopa and ropinirole, totaling a daily levodopa equivalent dose of 1,250 mg.
doc_15976	8	PD was in Hoehn-Yahr stage 3 while on medication, and the UPDRS-III score was 78 off medication and 18 after suprathreshold levodopa intake.

Рис. 3: (Figure:) Task-1 Cases/Sentences EN-ES Translation Examples: clinic-WMT21fb vs clinic-NLLB

term_n	**Transfer-learning**: clinical-WMT21fb:es2en	**Fine-tuning**: clinical-NLLB-200:es2en	Source: Spanish
term_1	Infantile paralysis	infantile paralysis	parálisis infantil
term_2	convulsive seizures	seizures	crisis convulsivas
term_5	deletion in chromosome 5 in the q15-q22 region	chromosome 5 deletion in the q15-q22 region	deleción en el cromosoma 5 en la región q15-q22
term_6	Familial adenomatous polyposis	familial adenomatous polyposis	poliposis adenomatosa familiar
term_9	Chromosomopathy	chromosomal disease	cromosomopatía
term_12	arterial hypertension	hypertension	hipertensión arterial
term_15	pT2bN0Mo clear cell renal adenocarcinoma	Renal clear cell adenocarcinoma pT2bN0Mo	adenocarcinoma renal de células claras pT2bN0Mo
term_17	hepatic lesions	liver lesions	lesiones hepáticas
term_18	Hepatic metastases	liver metastases	metástasis hepáticas
term_19	Metastatic renal cancer	metastatic renal cancer	cáncer renal metastásico
term_22	Deep vein thrombosis	deep vein thrombosis	trombosis venosa profunda
term_23	Asterixis	asterixis	asterixis
term_24	Aortic atheromatosis	aortic atheromatous disease	ateromatosis aórtica
term_29	hypothyroidism grade 2	grade 2 hypothyroidism	hipotiroidismo grado 2
term_30	Grade 3 hypertension	grade 3 hypertension	hipertensión arterial grado 3
term_31	Grade 3 diarrhea with secondary hypomagnesemia	grade 3 diarrhea with secondary hypomagnesemia	diarrea grado 3 con hipomagnesemia secundaria
term_32	Thrombocytopenia	thrombopenia	trombopenia
term_33	gastrointestinal toxicity	digestive toxicity	toxicidad digestiva
term_35	Recurrent respiratory tract infection	recurrent infectious respiratory	respiratoria infecciosa recurrente
term_36	Pulmonary nodule located in the upper lobe	pulmonary nodule located in the upper lobe	nódulo pulmonar localizado en el lóbulo superior
term_37	Loculated cystic lesion in LSD	Cystic lesion loculated in LSD	lesión quística loculada en LSD
term_38	Multicystic lesion	Multi-cystic lesion	lesión multiquística
term_43	MCVAP type I of LSD	LSD type I MCVAP	MCVAP tipo I del LSD
term_0	mild mental retardation	mild mental retardation	retraso mental leve
term_3	urinary tract infections	urinary tract infections	infecciones del tracto urinario
term_4	ITU) of repetition	ITU) of repetition	ITU) de repetición
term_7	deletion of this gene	deletion of this gene	deleción de dicho gen
term_8	deletion in chromosome 5	deletion in chromosome 5	deleción en el cromosoma 5
term_10	drug allergies	drug allergies	alergias medicamentosas
term_11	smoker	smoker	fumador
term_13	dyslipidemia	dyslipidemia	dislipemia
term_14	atrial fibrillation	atrial fibrillation	fibrilación auricular
term_16	macroscopic hematuria	macroscopic hematuria	hematuria macroscópica
term_20	hypothyroidism	hypothyroidism	hipotiroidismo
term_21	dehydration	dehydration	deshidratación
term_25	Cardiomegaly	Cardiomegaly	Cardiomegalia
term_26	anemia	anemia	anemia
term_27	hyponatremia secondary to diarrhea	hyponatremia secondary to diarrhea	hiponatremia secundaria al cuadro diarreico
term_28	sepsis	sepsis	sepsis
term_34	smoker	smoker	fumadora
term_39	cyst	cyst	quiste
term_40	microcytic anemia	microcytic anemia	anemia microcítica
term_41	ectopic pregnancy	ectopic pregnancy	embarazo ectópico
term_42	adenopathies	adenopathies	adenopatías

Рис. 4: (Figure:) Task-2 Clinical Term ES-EN Translation Examples: clinic-WMT21fb vs clinic-NLLB

the outputs of these two MMPLMs, such as on translating multi-word expressions in the clinical domain (Bhatia et al., 2023; Han, 2022). We also designed corresponding measurements on the evaluation of uncertainty and inter-rater reliability (IRR) levels (Gladkoff et al., 2022, 2023). Secondly, we think it is valuable to integrate more high-performance automatic metrics into the comparison such as hLEPOR (Han et al., 2021). Finally, we will try more external languages from different typologies in future work.

ont_n	**Transfer-learning**: clinical-WMT21fb (en2es)	**Fine-tuning**: clinical-NLLB (en2es)	Source:English
ont_0	Todos	Todos	All
ont_1	Anomalía de la altura corporal	Anomalías de la talla corporal	Abnormality of body height
ont_2	Displasia renal multiquística	Displasia renal multicística	Multicystic kidney dysplasia
ont_3	Displasia renal multiquística	Riñón displásico multicístico	Multicystic dysplastic kidney
ont_4	Riñón multiquístico	Riñones multicísticos	Multicystic kidneys
ont_5	Displasia renal multiquística	Displasia renal multicística	Multicystic renal dysplasia
ont_6	Modo de herencia	Modos de herencia	Mode of inheritance
ont_7	Herencia	Herencia	Inheritance
ont_8	Herencia autosómica dominante	Herencia autosómica dominante	Autosomal dominant inheritance
ont_9	autosómica dominante	Autosomal dominante	Autosomal dominant
ont_10	Forma autosómica dominante	Forma autosómica dominante	Autosomal dominant form
ont_11	Тип autosómico dominante	Tipo autosómico dominante	Autosomal dominant type
ont_12	Herencia autosómica recesiva	Herencia autosómica recesiva	Autosomal recessive inheritance
ont_13	autosómica recesiva	Autosomal recesivo	Autosomal recessive
ont_14	Forma autosómica recesiva	Forma autosómica recesiva	Autosomal recessive form
ont_15	Predisposición autosómica recesiva	Predisposición autosómica recesiva	Autosomal recessive predisposition
ont_16	Morfología anormal de los genitales internos femeninos	Morfología anormal de los genitales internos femeninos	Abnormal morphology of female internal genitalia
ont_17	Anomalía de los genitales internos femeninos	Anomalías de los genitales internos femeninos	Abnormality of female internal genitalia
ont_18	Anomalía funcional de la vejiga	Anomalías funcionales de la vejiga	Functional abnormality of the bladder
ont_19	Mal función vesical	Función vesical deficiente	Poor bladder function
ont_20	Infecciones urinarias de repetición	Infecciones urinarias recurrentes	Recurrent urinary tract infections
ont_21	Infecciones del tracto urinario frecuentes	Infecciones frecuentes del tracto urinario	Frequent urinary tract infections
ont_22	ITU recidivante	ITU recurrentes	Recurrent UTIs
ont_23	Infecciones vesicales de repetición	Infecciones vesiculares repetidas	Repeated bladder infections
ont_24	Infecciones urinarias de repetición	Infecciones urinarias repetidas	Repeated urinary tract infections
ont_25	Infecciones del tracto urinario	Infecciones del tracto urinario	Urinary tract infections
ont_26	Infecciones del tracto urinario, recurrentes	Infecciones del tracto urinario, recurrentes	Urinary tract infections, recurrent
ont_27	vejiga neurogénica	Vejícula neurogénica	Neurogenic bladder
ont_28	Falta de control vesical por lesión del sistema nervioso	Falta de control vesical por lesión del sistema nervioso	Lack of bladder control due to nervous system injury
ont_29	Urgencia urinaria	Urgencia urinaria	Urinary urgency
ont_30	vejiga hiperactiva	Vejícula hiperactiva	Overactive bladder
ont_31	Síndrome de vejiga hiperactiva	Síndrome de vejiga hiperactiva	Overactive bladder syndrome
ont_32	Síndrome de frecuencia de urgencia	Síndrome de frecuencia de urgencia	Urgency frequency syndrome
ont_33	Hipoplasia del útero	Hipoplasia del útero	Hypoplasia of the uterus
ont_34	Utero hipoplásico	Útero hipoplásico	Hypoplastic uterus
ont_35	Utero rudimentario	Útero rudimentario	Rudimentary uterus
ont_36	Utero pequeño	Útero pequeño	Small uterus
ont_37	Utero subdesarrollado	Útero subdesarrollado	Underdeveloped uterus
ont_38	Anomalía vesical	Anomalía vesical	Abnormality of the bladder
ont_39	Divertículo vesical	Divertículo vesical	Bladder diverticulum
ont_40	Divertículos vesicales	Divertículos vesiculares	Bladder diverticula
ont_41	Retención urinaria	Retención urinaria	Urinary retention
ont_42	Aumento del volumen residual de orina post-vacío	Aumento del volumen de orina residual post-vacío	Increased post-void residual urine volume
ont_43	La nicturia	Nocturia	Nocturia

Рис. 5: (Figure:) Task-3 Concept EN-ES Translation Examples: clinic-WMT21fb vs clinic-NLLB

Limitations

1) On PLM Capability for Transferring to New Language, in this work, we used Meta-AI's WMT21 multilingual pre-trained language models as our test-base for the knowledge transfer into an external language fine-tuning and translation. This new-language ability is much dependent on the MPLMs we used, such as WMT21fb (Tran et al., 2021) as a huge size model, a conditional generation from Meta-AI's massive M2M-100 model (Fan et al., 2021). If we try to fine-tune a bilingual model on an external language that the PLM did not see, it will not be that good because for smaller-sized models such fine-tuning would be too much of a change, and the model will lose generalisation which leads to problems. For huge multilingual PLM models, the 250K of fine-tuning data is a small set of numbers, and that's why the model does not lose generalisation and captures new data well without losing linguistic knowledge of other languages that it was trained on.

2) On the Impact of Language Families, the MMPLM WMT21fb we deployed has both alphabetic languages and CJK (Chinese, Japanese, Koran) character languages, as well as Slavic language (Russian). This might make it easier to transfer to a new language, e.g. alphabetic language. However, in situations when the MPLMs did not include any of the language scripts that belong to the language family of the target one, it can be much harder for it to transfer to the new target language. This needs further investigation. One possible extension of this work is using the dynamic vocabulary method proposed by Lakew et al. (2018).

Acknowledgements

The authors thank the ClinSpEn2022 shared task organisers for preparing the data set and evaluation platforms, and thank Darryl Estrada for communicating with us during the competition. We thank Dr. Alfredo Madrid Garcia and Ms. Cristina Sánchezfor carrying out the human evaluation for this work and thank anonymous reviewers for their valuable comments on our work. We thank the open research projects Meta-AI's wmt21.dense-24-wide.En-X (2021) and NLLB (2022) we used. This work has been partially supported by grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space"(funder: UKRI/EPSRC).

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACL.

Archna Bhatia, Kilian Evang, Marcos García, Voula Giouli, Lifeng Han, and Shiva Taslimipoor. 2023. Proceedings of the 19th workshop on multiword expressions (mwe 2023). In Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. J. Mach. Learn. Res., 22(1).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 866–875, San Diego, California. Association for Computational Linguistics.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 13–21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Lifeng Han, and Goran Nenadic. 2023. Student's t-distribution: On measuring the inter-rater reliability when the observations are scarce. arXiv preprint arXiv:2303.04526.

Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring uncertainty in translation quality evaluation (TQE). In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1454–1461, Marseille, France. European Language Resources Association.

Lifeng Han. 2022. An investigation into multi-word expressions in machine translation. Ph.D. thesis, Dublin City University.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022a. Examining large pre-trained language models for machine translation: What you don't know about it. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 908–919, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022b. Using massive multilingual pre-trained language models towards real zero-shot neural machine translation in clinical domain.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model LaBSE. In Proceedings of the Sixth Conference on Machine Translation, pages 1014–1023, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In Proceedings of the 15th International Conference on Spoken Language Translation, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 448–462, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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 of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. Technical report.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40, San Diego, California. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. CoRR, abs/2008.00401.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai's wmt21 news translation task submission. In Proc. of WMT.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Conference on Neural Information Processing System, pages 6000–6010.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. LREC MultilingualBIO: multilingual biomedical text processing.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.