LISN @ WMT 2021

Jitao Xu
Univ. Paris-Saclay,
& CNRS, LISN

Pham Minh Quang
Univ. Paris-Saclay,
& CNRS, LISN
& Systran

Sadaf Abdul Rauf
Univ. Paris-Saclay,
& CNRS, LISN

François Yvon
Univ. Paris-Saclay,
& CNRS, LISN

{firstname.lastname}@limsi.fr

Abstract

This paper describes LISN's submissions to two shared tasks at WMT'21. For the biomedical translation task, we have developed resource-heavy systems for the English-French language pair, using both out-of-domain and in-domain corpora. The target genre for this task (scientific abstracts) corresponds to texts that often have a standardized structure. Our systems attempt to take this structure into account using a hierarchical system of sentence-level tags. Translation systems were also prepared for the News task for the French-German language pair. The challenge was to perform unsupervised adaptation to the target domain (financial news). For this, we explored the potential of retrieval-based strategies, where sentences that are similar to test instances are used to prime the decoder.

1 Introduction

This paper describes LISN's[1] submissions to the translation shared tasks at WMT'21, where we took part in two shared tasks. For the biomedical translation tasks, we have developed resource-heavy systems for the English-French language pair, using a diversity of out-of-domain and in-domain corpora, thus continuing the efforts reported in (Abdul Rauf et al., 2020). Like for previous years shared task, the target genre (scientific abstract) corresponds to texts that often have a standardized structure comprising typical subsections of one to five lines. Standard subsections report the OBJECTIVE, the METHOD, or the RESULTs of the study. Our systems for this year attempt to take this structure into account using sentence-level tags, with the hope to capture some of the document structure and the phraseology of the domain into account. These systems are documented in Section 2.

Translation systems were also prepared for the News task for the French-German language pair. The challenge this year was to perform unsupervised adaptation to the target domain (financial news), with no further detail regarding the test data. In particular, the organizers did not release any development data to tune systems. In this setting, we explored the potential of using a retrieval-based strategy, where sentences that are similar to the test instances are used to help the decoding. In this approach, introduced in (Bulte and Tezcan, 2019) and further explored in (Xu et al., 2020; Pham et al., 2020), translation is a two-step process: a retrieval phase, which identifies sentences that resemble the source test sentence in parallel corpora. These sentences and their translation are then used to *prime* the decoder: inserting relevant translations examples in the decoder's context should help to select the right translations, especially for words and terms from the test domain. These systems are described in Section 3.

2 MT for biomedical texts

In this section, we describe our participation to the biomedical task for WMT'21, in which we participated in both English to French and French to English directions. English-French is a reasonably resourced language pair with respect to biomedical parallel corpora, allowing us to train our Neural Machine Translation (NMT) systems (Vaswani et al., 2017) with in-domain corpora as well as large out-of-domain data that exists for this language pair. Like for last year (Abdul Rauf et al., 2020), our first goal is to make the best of all the available data, including supplementary in-domain monolingual data. Our corpora are described in Section 2.1.

For this year's participation, we also attempt to take the internal structure of biomedical abstracts into account. Many of these abstracts follow what is often refered to as the "IMRAD format", comprising the following subparts: INTRODUCTION,

[1] LISN [Laboratoire Inderdisciplinaire des Sciences du Numérique] is the new name of the laboratory formerly known as LIMSI.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 232–242

Parallel			
Corpus	**Wrds (M)**		**Sents.**
	English	**French**	
Ufal	89.5	100.3	2.72 M
Edp	0.04	0.04	2.44 K
Medline titles	5.97	6.43	0.63 M
Medline abstracts	1.23	1.44	0.06 M
Scielo	0.17	0.21	7.84 K
Cochrane-Reference	2.23	2.74	0.12 M
Cochrane-PE	0.43	0.53	20.5 K
Cochrane-GooglePE	0.63	0.77	30.3 K
Taus	20.1	23.2	0.88 M
Mlia	19.0	23.0	1.0M
IR Retrieved	13.2	14.7	3.6M
Development			
Medline 18	5.7K	6.9K	265
Medline 19	9.8K	12.4K	537
Test			
Medline 20	12.7K	16.2K	699
Monolingual			
Corpus	**English**	**French**	**Sent.**
Lissa_Fr	8.79	7.70	0.33 M
Med_Fr	16.3	16.2	0.06 M
IsTex_Fr	6.92	7.84	0.42M
Med_En	3.40	4.02	0.22M
Out Domain			
Corpus	**English**	**French**	**Sent.**
Out-of-domain	1139	1292	35M

Table 1: Data sources for the biomedical task

METHODS, RESULTS, and DISCUSSION (Sollaci and Pereira, 2004). This structure can be explicit in documents through dedicated headings or remain implicit. Our experiments aim to explore how to use this information in NMT and to measure the correlated impact. We notably expect that by informing the system with sub-document information, it will learn the typical style and phraseology of sentences occurring in each part.

For this purpose, we identified in our data all the abstracts that were conforming to this basic structure and worked to make this structure as explicit and standardized as possible. This notably implied to normalize the mains headings, as some variation was observed: for instance, ANALYSIS may be replaced with DISCUSSION, and additional subparts (OBJECTIVES, CONCLUSION) are also be observed. To incorporate the standard IMRaD format we mapped each subheading to the corresponding IMRaD subpart using a system of tags. Details regarding this process are given in Section 2.2.

All systems are based on the Transformer architecture of Vaswani et al. (2017). We were able to achieve appreciable gains both from back-translation and document structure processing. The results are discussed in Section 2.4.

2.1 Corpus and preprocessing

We trained our baseline systems on a collection of in domain biomedical texts as well as out-of-domain parallel corpus. Table 1 details the corpora used in training.

2.1.1 Parallel corpora

We gathered parallel and monolingual corpora available for English-French in the biomedical domain. The former included the biomedical texts provided by the WMT'20 organizers: Edp, Medline abstracts and titles (Jimeno Yepes et al., 2017), Scielo (Neves et al., 2016) and the Ufal Medical corpus[2] consisting of Cesta, Ecdc, Emea (OpenSubtitles), PatTR Medical and Subtitles. In addition, we used the Cochrane bilingual parallel corpus (Ive et al., 2016)[3], the Taus Corona Crisis corpus[4] and the Mlia Covid corpus.[5] We finally experimented with additional in-domain data selected using Information Retrieval (IR) techniques from general domain corpora including News-Commentary, Books and Wikipedia corpus obtained from the Open Parallel Corpus (OPUS) (Lison and Tiedemann, 2016). These were selected using the data selection scheme described in (Abdul-Rauf and Schwenk, 2009). Medline titles were used as queries to find relevant sentences. We used the `2-best` sentences returned from the IR pipeline as additional corpus.

Our out-of-domain corpora include the parallel data provided by the WMT14 campaign for French-English: Gigafr-en, Common Crawl, Europarl, News Commentary and the UN corpora.

For development purposes, we used Medline test sets of WMT'18 and 19, while Medline 20 was used as internal test data.[6]

[2] https://ufal.mff.cuni.cz/ufal_medical_corpus

[3] https://github.com/fyvo/CochraneTranslations/

[4] https://md.taus.net/corona

[5] http://eval.covid19-mlia.eu/task3/

[6] These testsets were sentence-aligned with in-house

2.1.2 Monolingual sources

The back-translation of monolingual sources has often been effectively used to cater for parallel corpus shortage in the Biomedical domain in (Stojanovski et al., 2019; Peng et al., 2019). We also adopt this approach here.

Supplementary French data from three monolingual sources were collected from public archives: abstracts of medical papers published by Elsevier from the Lissa portal[7] and from the national ISTEX archive[8]; a collection of research articles collected from various sources[9] henceforth referred to as Med_Fr (Maniez, 2009). These documents were automatically translated into English with an NMT system trained on biomedical corpora, with a BLEU score of 33.6 on Medline20 testset.

The English side of Medline German and Spanish corpora is used as supplementary English data for back translation. Duplicate documents were removed based on the document id. For these, the internal structure of documents is often available and has been tagged as for the parallel data. These texts were then split into sentences[10] and translated into French using a NMT system trained on all Biomedical corpora with a BLEU score of 36.4 on Medline20 testset. All back-translated data is tagged using the proposal of Caswell et al. (2019).

Parallel and monolingual data are further processed using SentencePiece (Kudo and Richardson, 2018) tokenisation and detokenisation scheme to segment texts into subword units using a vocabulary of 32K subwords. These units were learned on all the in-domain corpora.

2.2 Sentence tagging: a three-level scheme

2.2.1 Tagging domains and corpora

As explained above, our training data is diverse, comprising in-domain parallel, out-of-domain parallel, and in-domain monolingual that is automatically back-translated. Some are made of lists of isolated sentences, while others retain the document information. Even within the in-domain data, some texts precisely match the genre of the testset (scientific abstracts) - this is the case for instance of Medline and to a lesser extent, Cochrane; while others are more remote (eg. the Ufal collection, or the Mlia corpus). In order to reflect this diversity, we designed a three-level sentence tagging scheme that is used for the experiments in Section 2.4.2. These tags appear as prefix of each source sentence.

The first level of tags distinguishes between out-of-domain data (<G>), and in-domain data (tagged <M>). The second level of tag aims to distinguish between data sources, hence the use of one dedicated tag for each corpus, except for the monolingual data, which is simply tagged with <BT>.

2.2.2 Tagging sections within documents

The third level of annotation is indented to enhance the translation context with information regarding the position of a sentence within the abstract. The structure of scientific abstracts in the medical domain often obey the IMRAD structure, and the third tag aims to include this structural information as an additional document-level context. Document level information is necessary to model long-range dependencies between words, phrases, or sentences, or document parts. For a translation system, the ability to model the context may notably improve certain translation decisions, e.g. a better or most consistent lexical choice (Kuang et al., 2018) or a better translation of anaphoric pronouns (Voita et al., 2018; Bawden et al., 2019). A recent review of these themes is in (Maruf et al., 2021).

For this purpose, we further pre-processed 6 corpora containing scientific abstracts. These corpora had different subheadings and structures as given below, which were mapped to a restricted set of section tags listed in Table 2:

1. Medline and Scielo: Abstracts and sub headings often without title. We identified a total of 189 subheadings including spelling variations. Examples include: Presenting Concerns of the Patient, Sources of Information, Novel finding, Study Selection etc.

2. Edp: Abstracts and sub headings mostly contain titles. 45 subheadings where found, such as: Case report, Observation, Subjects and Methods, Commentary, Pedagogical objectives etc.

3. Cochrane: only 10 different subheadings were found, including: Abs selection criteria, abs search strategy, abs data collection, summary title etc.

The identification and standardization of subheading information was a tedious process, involving a lot of rule-based processed to take the variability of sub-headings into account. In order to reconstruct fully parallel versions with subheadings, we also had to reinsert explicit headings in

tools and are shared at https://github.com/fyvo/WMT-Biomed-Test.

[7] https://www.lissa.fr/dc/#env=lissa

[8] https://www.istex.fr/

[9] https://crtt.univ-lyon2.fr/les-corpus-medicaux-du-crtt-613310.kjsp

[10] https://pypi.org/project/sentence-splitter/

Title	<H1>
Introduction	<INT>
Objectives	<OBJ>
Material and Methods	<MaM>
Results	<RES>
Conclusion	<CON>

Table 2: Standardized section heading tags

the source or the target files. Also note that this information was not available for all abstracts. After preprocessing files for which the full subheading information was available, we obtained the 6 fully-tagged corpora (see statistics in Table 3). A similar process was used for test sets (see Table 4).

Corpus	Lines	En words	Fr words
Medline	34836	742891	920811
Edp	1682	34167	37508
Scielo (wmt16)	7088	163275	199829
Cochrane-Reference	123598	2741426	3308485
Cochrane-GooglePE	30866	685490	828436
Cochrane-PE	20693	468691	568262

Table 3: Document-aligned training corpora

Testset	en-fr	fr-en
medline20	735	580
medline18	321	347
medline19	493	469

Table 4: Number of test sentences after alignment

Finally, we also introduced a third tag in all other documents as follows: sentences within an abstract where tagged as <ABS>, while all remaining sentences from other corpora where simply tagged as "unspecified subheading" (<US>).

2.3 Translation framework

Our translation systems mostly used the basic Transformer models, while a few contrastive systems used the large version (Vaswani et al., 2017). They all rely on Facebook's seq-2-seq library (fairseq) (Ott et al., 2019) with parameters settings borrowed from `transformer_wmt_de_en`.[11]. The ReLU activation function was used in all encoder and decoder layers. We optimize with Adam (Kingma and Ba, 2015), set up with a maximum learning rate of 0.0005 and an inverse square root decay schedule, as well as 4000 warmup updates. We share the decoder input and output embedding matrices. Models are trained with mixed precision and a batch size of 4096 tokens on 4 V100 GPUs for 300k updates. Systems were trained until convergence based on the BLEU score on the development sets. Evaluation was performed using SacreBleu (Post, 2018). Scores are chosen based on the best score on the development set (Medline 18, 19) and the corresponding scores for that checkpoint are reported on Medline 20 test set.

For fine-tuned systems, the process starts with models trained to convergence, based on BLEU score on dev sets. Training then resumes using a selected portion of the training corpus using the same parameters and criterion as for the base systems. In our results corresponding systems are post-fixed with `*-ft`.

2.4 Results

We present our results for the two directions in two tables, Table 5 and 6, differentiating the normal versus the *tag-based* systems. Base systems are given on the left, ($\Rightarrow$) identifies the derived (fine-tuned) systems.

2.4.1 Regular MT systems

Results for the untagged systems are reported in Table 5 and are denoted by `X*`, with `E*` and `F*` representing the English to French and French to English systems respectively.

We first built baseline systems. `X0` denotes the systems built using only the in-domain data provided by the organizers. `X1` are our baseline systems built using all in-domain parallel data. We see good improvement in both directions amounting to 4.2 and 4.8 BLEU points, which is obtained by adding around $1M$ sentences of additional Cochrane and Taus corpora to the already available $3.4M$ sentences from WMT'20. This hints at the relevance of the additional in-domain parallel corpora used.

We used the `X1` systems as strong in domain baselines to study the effect of adding back-translated in domain data. These appear as `X2` and `X3` in Table 5. Adding around 0.8M French to English and around 0.2M English to French back translated sentences did not help as much as we were expecting. We saw similar results last year and increased the amount of back translations this

[11] https://fairseq.readthedocs.io/en/latest/models.html

ID	Train	ID	Sentences	Medline 20	ID	Sentences	Medline 20
			EN-FR			**FR-EN**	
X0	WMT biomed data	**E0**	3.4M	31.6	**F0**	3.4M	28.8
X1	All biomed	**E1**	4.5M	35.8	**F0**	4.5M	33.6
Back translations of monolingual data							
X2	X1 + BT	**E2**	5.3M	34.8	**E2**	4.7M	33.5
X3	X1 + BT-tag	**E3**	5.3M	**36.6**	**F3**	4.7M	32.4
Out of domain fine-tuned with in domain							
X4	outdomain⇒biomed	**E4**	40.5M	32.3	**F4**[3]	41M	**35.8**

Table 5: Results for systems using in-domain and out-of-domain corpora. Superscripts ^{*n} denote runs submitted

ID	Train	Sentences	Medline 20		
			<SUBHEAD>	<ABS>	<US>
			EN-FR		
TE1	Out+In	41.7M	36.2	**36.3**	**36.3**
TE2[1]	TE1⇒ftbiomedplusbt	47.2M	**38.7**	38.5	38.6
TE3	TE2⇒ftCocMed	48.0M	38.2	**38.4**	38.3
Transformer Large					
TE4	Out+In	41.7M	36.1	36.2	**36.3**
TE5[2]	TE4⇒ftbiomedplusbt	47.2M	38.4	**38.5**	38.2
			FR-EN		
TF1	Out+In	40.9M	**32.1**	32.0	**32.1**
Mixed baseline finetuned with in-domain					
TF2[1]	TF1⇒ftbiomedplusbt	46.4M	**35.7**	35.2	35.2
TF3[2]	TF2⇒ftCocMed	48.8M	**35.3**	34.9	34.8

Table 6: Results for systems with sentences tagged with our 3 level tagging scheme. Test sets are decoded 3 times, where the third tag is varied from the more specific (<SUBHEAD>) to the more generic (<US>). Superscripts ^{*n} denote the runs submitted.

year. `X3` denote systems built using the tagging scheme proposed by Caswell et al. (2019), where back translations are prefixed with the <BT> tag on the source side.

Indicating that a training sentence is back-translated allows the model to separate the helpful and harmful signal. This proved particularly true for English into French where adding tag to back translations improved the BLEU score by 0.8 points; but it was not helpful in the reverse direction where the amount of back translated data was may be too small (0.2M lines). back-translations as compared to the baseline corpora.

Finally, systems were built by initialising the parameters from huge out-of-domain corpora and later fine tuned on in-domain corpora (`X4`), where in-domain sub words learned from all the Biomedical data are used to segment the out-of-domain data. The initial systems were trained for 4 epochs on general domain WMT14 EN-FR corpora. The FR-EN system (`F4`) is the best system in this direction, reaching a BLEU score of 35.8.

2.4.2 Tagged Systems

As our 3-level tagging scheme, described in Section 2.2, is adding information about the domain of each sentence, we specifically focused on larger systems by using all the available in- and out-of-

domain corpora.

Results are summarized in Table 6 with TE* representing the Tagged English to French systems and TF* representing the French to English systems. TE1 is the baseline system for EN-FR built with all the available in domain and out-domain data. TE4 is the corresponding baseline using a Large Transformer[12]. We then fine-tune these systems with all the in-domain data including the back translations, these are represented by TE2 and TE5 respectively. This gives an appreciable gain of 2.5 and 2.3 BLEU points for Transformer and Transformer large systems. As we saw no major difference in scores for Transformer versus Transformer large, so we continue the rest of experiments with the simple Transformer architecture. Fine-tuning further with just abstracts from Cochrane and Medline did not yield any further improvement.

French to English results display similar trends. The baseline (TF1) using all available (in domain + out-of-domain) data tagged with our 3 level scheme yielded a BLEU score of 32.1. Fine-tuning it further with all in-domain data (TF2) gives an improvement of 3.6 BLEU points which does not improve further when fine-tuning continues with just Cochrane and Medline abstracts (TF3).

To measure whether the model learned document domain and/or sentence origin information, we tested by tagging the test set with three different tags in the third position, using either the exact subheading, or abstract or UnSpecified for sentences for which the sub-section is unknown. Table 6 reports the scores for the three cases. Though the difference in scores for the three cases is minute, in-domain systems{TE2, TE3, TE5} and {TF2, TF3} achieve their best results when the test set is tagged with the subheading or the abstract tag, typical feature of the biomedical corpora. Conversely, for out-of-domain systems {TE1, TE4, TF1}, the best scores are always for the test set tagged with <US>. This strongly hints that the system is using the extra-information provided by the tag. These observations need to be confirmed using other metrics, as BLEU may not properly reflect these differences.

For English to French direction we got better scores with the tagged systems, with the best system (TE2 = 38.7) achieving 2.1 BLEU points more than the best un-tagged system (E3 = 36.6). This was however not the case for French-English where both tagged and un-tagged systems had more or less similar scores.

[12]hidden size of 1024 and a feed forward size of 4096. Rest of the parameters same as for other systems.

		EN-FR		
E2	base+bt	34.8		
E3	base+bt-tag	36.6		
		<SUBHEAD>	<ABS>	<US>
TE	Indomain+bt	**37.3**	37.0	37.0
		FR-EN		
F2	base+bt	33.5		
F3	base+bt-tag	32.4		
		<SUBHEAD>	<ABS>	<US>
TF	Indomain+bt	**34.4**	34.4	34.4

Table 7: Comparison of our 3 level tagged systems with the corresponding untagged systems. Systems {E2,F2} are built by adding back-translated data to the baseline. In systems {E3,E3}, the added back-translated data start with <BT> tag. Systems {TE, TF} use our 3-level tagging scheme for all sentences.

Systems in Tables 5 and 6 have different baselines, thus to establish a fair comparison we report numbers for comparable systems in Table 7. Systems {E2,E3,F2,E3} are copied from Table 5, whereas {TE and TF} are the corresponding systems using our tagging scheme on the sole biomedical data. We see here a clear gain for French-English when we use a 3-level tagging scheme (TF) compared to just adding the <BT> tag (F3); results for the reverse direction are more even and having one or three tags does not make a difference.

2.5 Conclusion

In this section, we have presented our work for the biomedical task. We notably have tried to incorporate document origin and structure information and improve strong baseline systems that were using a wealth of in-domain and out-of-domain data. . Overall, our systems for this year are significantly better than last year's, even though the benefits of adding document structures as tags need to be confirmed by more experiments and analyses.

3 News translation task: De ↔ Fr

In the 2021 News translation task, we focused on the German-French language pair in which the participants are asked to build MT systems for News in the financial domain. In this section, we discuss details of our approach and the rationale behind it.

3.1 Unsupervised adaptation

As the training and development data do not contain domain information, the supervised domain adaptation paradigm is not suitable here. However, non-parametric adaptation (Bapna and Firat, 2019), example-based guided machine translation (Zhang et al., 2018), unsupervised domain adaptation (Farajian et al., 2017) or priming NMT (Xu et al., 2020; Pham et al., 2020) have showed promising results for this problem. These approaches retrieve translation examples that are similar to the input source sentence, and use them to guide the inference and to reproduce existing translations or to locally adapt the pre-trained NMT system to the input sentence.

Even though all of the approaches mentioned above have merits of their own, we decided to focus on computationally cheaper methods such as (Bulte and Tezcan, 2019; Xu et al., 2020) where the retrieved instances provide an extra conditioning context for the decoder. Pham et al. (2020) further improved these techniques by proposing to simultaneously prime the source and the target side of the retrieved examples (see Section 3.4.1) and has been our main source of inspiration.

3.2 Data and preprocessing

We use all available parallel data for De $\leftrightarrow$ Fr, with the exception of the ParaCrawl data, for training. We also use monolingual data to improve translation quality. For both languages, we choose Newscrawl 2020. We additionally use Newscrawl 2018 and 2019 French data at inference time to explore the ability of our priming model to make use of extra data. Details are in Section 3.4.2. We use *newstest2019* as development set and test our models on *newstest2020*.

We filter out sentence pairs with invalid language tag using `fasttext` language id model[13] (Bojanowski et al., 2017). We use Moses tools to normalize punctuation, to remove non-printing characters and to tokenize into words. The final parallel data contains 5.6M sentences.[14] We use a shared source-target vocabulary built with 40K Byte Pair Encoding (BPE) units using the `subword-nmt` implementation (Sennrich et al., 2016b).[15]

[13] https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin

[14] https://github.com/moses-smt/mosesdecoder

[15] https://github.com/rsennrich/subword-nmt

3.3 Baseline systems

We build our Transformer-based (Vaswani et al., 2017) systems using `fairseq`[16] (Ott et al., 2019). Our baseline system is a large Transformer with a hidden size of 1024 and a feedforward size of 4096. We optimize with Adam (Kingma and Ba, 2015), set up with a maximum learning rate of 0.0007 and an inverse square root decay schedule, as well as 4000 warmup updates. We tie the encoder and decoder input embedding matrices with the decoder output embedding matrix and we apply layer normalization before each block. Models are trained with mixed precision and a batch size of 4096 tokens on 4 V100 GPUs for 300k updates.

3.4 Submitted systems

3.4.1 Boosting NMT by similar translations

Our approach comprises 2 steps: similar translation retrieval and inference where the priming example is processed in forced-decoding mode.

The retrieval of relevant examples for a given source sentence is based on their distance in some high-dimensional numerical representation space. These representations are computed using the encoder of the baseline system (see Section3.3) so as to keep our systems in the "constrained" track, as the use of large pre-trained models such as BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), etc., was only allowed in unconstrained submissions. More precisely, for each sentence, we average the contextualized embeddings output at the last layer of the encoder. From the training dataset, we create a datastore of pairs (K, V) in which the key K is the sentence embedding of some source sentence $\mathbf{f}$ and the value is the sentence pair $(\mathbf{f}, \mathbf{e})$ whose source sentence is $\mathbf{f}$. For each query, we retrieve k keys ($k = 10$ in all experiments).

The similarity between two sentences is the cosine similarity and the retrieval of the nearest neighbor(s) is performed using FAISS (Johnson et al., 2017). In order to search through a large datastore, we divide it into shards containing at most 500K data points; we conduct the k nearest neighbor search on each shard, gather all the retrieved keys from all shards into a list and reduce it to the k nearest keys. Given an input sentence and the list of its k nearest neighbours, we append m ($m \leq k$) retrieved source sentences to the input sentence and initialize the target side by the concatenation of the

[16] https://github.com/pytorch/fairseq

m corresponding target sentences. We use a special token to separate sentences.

During training, we train the NMT model with two types of examples (with and without retrieval): this means that each training sample will occur twice, once with and once without priming. The former examples have the following format:

$$\mathbf{f}_1 * \cdots * \mathbf{f}_m || \mathbf{f}$$

$$\mathbf{e}_1 * \cdots * \mathbf{e}_m || \mathbf{e}$$

while the latter is presented as the original data.

During inference, we use the same format as for the source-side, while we initialize the decoder with the prefix $\mathbf{e}_1 * \cdots * \mathbf{e}_m ||$. We therefore call this initialization "force-decoding". The special tokens, which serve as joiners between the retrieved sentences and the source/target sentence, are carefully chosen so that they never occur in the real text to avoid ambiguity. As discussed in Pham et al. (2020), it is possible to concatenate several similar sentences i.e. use $m > 1$; we however only report results with $m = 1$, since our preliminary experiments did not show superior results with $m > 1$.

3.4.2 Monolingual retrieval

Pham et al. (2020) suggested that monolingual texts in the target language can also be helpful to inform the inference. To make use of monolingual data, we create pseudo translation pairs with back-translation to generate the missing source language side. For this step, we leverage the baseline NMT system in Section 3.3 for one direction to back-translate the monolingual target text in the inverse direction. We use Newscrawl 2020 as monolingual resource for both directions. The monolingual French data contains approximately 10M sentences while the German data is much larger. We randomly extract 10M sentences from the German monolingual data as the pseudo corpus. The back-translated corpora are added to the real parallel corpora to create a larger datastore for retrieval.

3.5 Evaluation

3.5.1 Priming and Back-translation

We mainly evaluate our method on the De→Fr direction. Results on both Newstest2019 and 2020 are in Table 8. Our priming model is able to improve for 0.4 BLEU on newstest2019. However, the same improvement is not observed for newstest2020. As indicated in Pham et al. (2020), monolingual back-translated data could be directly applied during inference without any additional training. We thus search similar sentences on both original and synthetic data for the test sets. As shown in Table 8 (+ bt inference), searching on synthetic data directly improves our results by 0.6 BLEU point on newstest2019.

Model	newstest2019	newstest2020
baseline	35.7	32.8
+ bt	37.5	33.7
+ tag	37.5	34.3
priming	34.6	33.2
+ bt inference	35.2	33.2
priming + bt	37.4	33.9
+ tag	36.9	34.1
+ min sim 0.85	37.5	34.3

Table 8: BLEU scores of models for De→Fr direction. Our best submitted system obtained a BLEU score of 28.1 on newstest2021.

Even though priming model could benefit from back-translated data at inference time, training with synthetic data has proven to be effective in many previous works (Sennrich et al., 2016a; Edunov et al., 2018; Ng et al., 2019). Therefore, we also experiment by adding back-translated data to the original data and retrain a translation model. Results (+ bt) demonstrate that training with synthetic data clearly improves the performance on both test sets. Caswell et al. (2019) reports that using explicit tags to distinguish original from back-translated data provides further gains; however in our experiments, tagging BT data was not very helpful.

Our model using priming with synthetic data was not able to surpass the baseline model trained with additional back-translated data. One possible reason is that similar sentences retrieved with low similarity scores may be too noisy, and therefore decrease the overall performance. Filtering out noisy similar sentences (with a threshold of 0.85)[17] help to further improve the performance and makes it our best system (+ min sim 0.85). This setting was used for our primary submission for both directions.

We directly apply the best settings found for De→Fr to the reverse direction (Fr→De) and report the corresponding results in Table 9.

[17]Thresholding the minimum similarity score is the result of a trade-off: using a high threshold selects good sentences for priming, at the risk of leaving many examples without any priming data, while a low threshold retrieves more examples, many of which are of poor quality. Our preliminary experiments showed that that 0.85 was a reasonable value.

Model	newstest2019	newstest2020
baseline	27.7	27.2
+ bt	32.4	32.9
+ tag	30.9	31.0
priming + bt	29.8	29.3
+ tag	29.5	29.6
+ min sim 0.85	30.4	30.1

Table 9: BLEU scores of models for Fr→De. Our best submitted system obtained a BLEU score of 37.2 on newstest2021.

3.5.2 Priming and domain adaptation

In this section, we try to assess the relationship between domain adaptation (DA) and priming, and question our initial assumption that priming performs some kind of unsupervised adaptation. Our test set for this part contains 1000 lines extracted from the European Central Bank (ECB) corpus, also available from OPUS website.

As an alternative to priming, we first consider a simple unsupervised domain adaptation technique, where we retrieve $k = 10$ most similar sentences for each test sample, yielding a corpus of $10 \times k$ sentences that we use to fine-tune for two epochs the baseline systems. Again, filtering based on a similarity scores helps to accumulate a smaller number of sentences that are closer to the test domain.

We then try to combine priming and fine-tuning in the following manner: for each test sentence, we use the k nearest examples $(\mathbf{f}_1, \mathbf{e}_1) \ldots (\mathbf{f}_k, \mathbf{e}_k)$ to derive k *domain-adaptation examples with priming* as follows: the first primes $\mathbf{f}_2$ with $\mathbf{f}_1$, the second $\mathbf{f}_3$ with $\mathbf{f}_2$, and so on, until finally $\mathbf{f}_1$ is primed with $\mathbf{f}_k$ (the target part is built accordingly). This corpus is used for fine-tuning, and decoding proceeds as before (with $\mathbf{f}_1$ as prime).

These approaches (priming, unsupervised DA, and priming+DA) are compared in Table 10. We first see that using back-translated data is detrimental to the BLEU score of the baseline system, an effect that might be due to the difference between News texts and ECB domain. We also see that unsupervised adaptation with highly similar sentences yields a small gain. Priming alone achieves the same result as the baseline, but can also benefit somewhat from unsupervised DA. Our best results are obtained when we mix the two strategies, only keeping highly similar sentences.

Model	ECB
baseline	26.7
baseline + bt + tag	25.9
+ FT min sim 0.7	26.3
+ FT min sim 0.8	26.1
priming + bt + tag	25.9
+ FT	25.6
+ FT min sim 0.7	26.3
+ FT min sim 0.8	26.0
priming + bt + tag + min sim 0.7	26.3
+ FT min sim 0.7	26.5
priming + bt + tag + min sim 0.8	26.3
+ FT min sim 0.8	26.3

Table 10: BLEU scores for De→Fr on ECB.

3.6 Conclusion

In this section, we have reported our attempt to perform domain adaptation through priming, a technique which uses sentences that are similar to the test instances to provide additional context in training and decoding. In our experiments with the translation of News between French and German, we had little success with this technique, even when using massive amounts of back-translated data to search for relevant primes. This suggests that priming is not so useful for "open" domains such as News (Pham et al., 2020), and should better be used for standardized types of texts that occur in more specialized domains. We also tried to compare unsupervised DA and priming, showing that, in our context, the former was yielding better results that than the latter and also proposed a promising way to combine these two complementary techniques.

4 Conclusion and outlook

In this paper, we have described the systems prepared for this year's participation to WMT shared tasks. For the biomedical track, most of our efforts have been invested in the development of high resource systems, trying to take the structure of medical abstracts into account. In the News task, we have explored ways to perform unsupervised domain adaptation using retrieval based techniques and back-translated data.

Acknowledgements

This work was made possible thanks to the Saclay-IA and the Jean ZAY computing platforms. It was granted access to the HPC resources of IDRIS under the allocation 2021-[AD011011580R1, AD011011270R1, AD011011717] made by

GENCI. The first author is funded through a regional grant from the “Région Ile de France”.

References

Sadaf Abdul Rauf, José Carlos Rosales Núñez, Minh Quang Pham, and François Yvon. 2020. LIMSI @ WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 803–812, Online. Association for Computational Linguistics.

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. 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 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Julia Ive, Aurélien Max, François Yvon, and Philippe Ravaud. 2016. Diagnosing high-quality statistical machine translation using traces of post-edition operations. In *International Conference on Language Resources and Evaluation - Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016 2016)*, page 8, Portorož, Slovenia.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

François Maniez. 2009. L'adjectif dénominal en langue de spécialité: étude du domaine de la médecine. *Revue française de linguistique appliquée*, 14(2):117–130.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The Scielo Corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Peng, Jianfeng Liu, Liangyou Li, and Qun Liu. 2019. Huawei's NMT systems for the WMT 2019 biomedical translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, Florence, Italy. Association for Computational Linguistics.

Minh Quang Pham, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. 2020. Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527, Online. 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, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Luciana B Sollaci and Mauricio G Pereira. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association : JMLA*, 92(3):364–367.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2019. The LMU munich unsupervised machine translation system for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 393–399, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.