LIMSI @ WMT 2020

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

José Carlos Rosales
Univ. Paris-Saclay,
& CNRS, LIMSI
& Inria Paris

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

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

{firstname.lastname}@limsi.fr

Abstract

This paper describes LIMSI's submissions to the translation shared tasks at WMT'20. This year we have focused our efforts on the biomedical translation task, developing a resource-heavy system for the translation of medical abstracts from English into French, using back-translated texts, terminological resources as well as multiple pre-processing pipelines, including pre-trained representations. Systems were also prepared for the robustness task for translating from English into German; for this large-scale task we developed multi-domain, noise-robust, translation systems aim to handle the two test conditions: zero-shot and few-shot domain adaptation.

1 Introduction

This paper describes LIMSI's submissions to the translation shared tasks at WMT'20. This year we have focused our efforts on the biomedical translation task, developing a resource-heavy system for the translation of medical abstract from English into French, using back-translated texts, terminological resources as well as multiple pre-processing pipelines, including pre-trained representations. Systems where also prepared for the robustness task for translating from English into German; for this large-scale task we developed multi-domain, noise-robust, translation systems aim to handle the two test conditions: zero-shot and few-shot domain adaptation.

Machine translation for the biomedical domain is gaining interest owing to the unequivocal significance of medical scientific texts. The vast majority of these texts are published in English and Biomedical MT aims to also make them available in multiple languages. This is a rather challenging task, due to the scope of this domain, and the corresponding large and open vocabulary, including terms and non-lexical forms (for dates, biomedical entities, measures, etc). The quality of the resulting MT output thus varies depending on the amount of biomedical (in-domain) resources available for each target language.

We participated in this years WMT'20 biomedical translation evaluation for English to French direction. English-French is a reasonably resourced language pair with respect to Biomedical parallel corpora, allowing us to train our Neural Machine Translation (NMT) (Sutskever et al., 2014) with only in-domain corpora and dispense with the processing of large out-of-domain data that exist for this language pair. Our main focus for this year's participation was to develop strong baselines by making the best of auxiliary resources: back translation of monolingual data; partial pre-translation of terms; pre-trained multilingual contextual embeddings and IR retrieved in domain corpora. Two pre-prossessing pipelines, one using the standard Moses tools[1] and subword-nmt (Sennrich et al., 2016b) and other using HuggingFace BERT API were developed and compared. All systems are based on the transformer architecture (Vaswani et al., 2017), or and on the related BERT-fused transformer model of Zhu et al. (2020). If our baselines were actually strong, we only managed to get relatively small gains from our auxiliary resources, for reasons that by and large remain to be analyzed in depth. Our biomedical systems are presented in Section 2.

We also participated in the Robustness translation task, developing a multi-domain, noise-robust and amenable to fast adaptation translation system for the translation direction English-German. Our main focus was to study in more depth the adaptor architecture initially introduced in (Bapna and Firat, 2019) in a large-scale setting, where multiple heterogeneous corpora of unbalanced size are available for training, and explore ways to make the system robust to spelling noise in the test data. The zero-shot system is a generic system which

[1] http://www.statmt.org/moses/

Proceedings of the 5th Conference on Machine Translation (WMT), pages 803–812

does not use any adaptation layer; for our few-shot adaptation submission, we did not use the supplementary data provided by the organizers, which turned out to be only mildly relevant for the test condition, but resorted to a data selection strategy. In any case, our submissions are constrained and only use the parallel WMT data for this language pair; they are further described in Section 3.

2 Bio-medical translation from English into French

2.1 Data sources

We trained our baseline systems on a collection of biomedical corpora, *excluding by principle any out-of-domain* parallel corpus, so as to keep the size of our systems moderate and a reduced training time. Table 1 details the corpora used in training.

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	8.86 M
IR Retrieved	13.2	14.7	3.6M
Development			
Scielo	0.09	0.13	4333
Edp	6.2K	7.1K	328
Khresmoi	28K	33K	1500
Test			
Medline 18	5.7K	6.9K	265
Medline 19	9.8K	12.4K	537
Medline 20	12.7K	16.2K	699
Monolingual			
Corpus	**English** *(Synthetic)*	**French** *(Human)*	**Sent.**
Lissa	8.79	7.70	0.33 M
Med_Fr	16.3	16.2	0.06 M

Table 1: Data sources for the English-French biomedical task (before tokenization)

We gathered parallel and monolingual corpora available for English-French in the biomedical domain. These first 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 OpenSubtitles. In addition, we used the Cochrane bilingual parallel corpus (Ive et al., 2016)[3] and the Taus Corona Crisis corpus.[4]. 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 Open Parallel Corpus (OPUS) (Tiedemann, 2012). These were selected using the data selection scheme described in (Abdul-Rauf and Schwenk, 2009). Medline titles were used as queries to find the related sentences. We used `3-best` sentences returned from the IR pipeline as additional corpus to build the models (these are shown as X7 in table2).

For development purposes, we used Khresmoi, Edp and Scielo test corpora. The Medline test sets of WMT'18 and 19[5] were used as internal test data.

2.1.1 Monolingual sources

Supplementary French data from two monolingual sources were collected from public archives: abstracts of medical papers published by Elsevier from the Lissa portal[6] and a collection of research articles collected from various sources[7] henceforth referred to as Med_Fr (Maniez, 2009). The former corpus contains 41K abstract and totals approximately 7.7M running words; the latter contains 65K sentences, for a little more than 1.5M running words.

These texts were back-translated (Sennrich et al., 2016a; Burlot and Yvon, 2018) into French using a relatively basic neural French-English engine trained with the official WMT data sources for the biomedical task, using the HuggingFace pipeline (see details below). This system had a BLEU score of 31.2 on Medline 18 test set.

Note that back-translation has also been effec-

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

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

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

[5] With our own sentence alignment.

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

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

Symptoms of bacterial pneumonia frequently overlap those present with viral infections or reactive airway disease.
Symptoms of pneumonie bactérienne frequently overlap those present with infections virales or reactive airway maladie.

Figure 1: An example sentence containing pre-translated terms in French

tively used to cater for parallel corpus shortage in the Biomedical domain in (Stojanovski et al., 2019; Peng et al., 2019; Soares and Krallinger, 2019).

2.2 Pre and post-processing

The document level corpora were first retrieved from xml, split[8] into sentences and sentence aligned using Microsoft bilingual aligner (Moore, 2002): these include Cochrane, Scielo and some unaligned documents from Edp. All train, development and test corpora were cleaned by removing instances of empty lines, URLs and lines containing more than 60% non-alphabetic forms.

For tokenization into words and subwords units, two pipelines were considered. The first one is set up as follows (a) tokenize the French and English texts using Moses scripts[9]; (b) compute a joint Byte-pair Encoding (BPE) inventory of 32K units with subword-nmt;[10] (c) generate the translation; (d) detokenize and truecase the output, again with Moses scripts. Systems based on this pipeline are prefixed `M*`. The second one is slightly more complex as it heavily relies on the HuggingFace API[11] for accessing pre-trained BERT models. The corresponding systems are prefixed with `H*` and comprise the following steps: (a) a simple tokenization script, (b) a multilingual segmenter mapping BPE units to pre-trained encodings generated according to (Devlin et al., 2019) as input to the translation system (step (c)). In that case, the MT output is also a sequence of multilingual BPE units that further needs (d) to be reaccentuated and recased, before a final (e) detokenization. Step (d) is non-trivial and is performed by a monolingual translation system trained to convert HuggingFace BPE units into Moses BPE units,[12] which can then be properly reassembled and detokenized as for the Moses pipeline.

2.2.1 Fine-tuning

The fine-tuning process starts from corresponding models trained to convergence, based on BLEU score on dev sets. These are then further fine-tuned using a selected part of the training corpus containing only the Medline abstracts and the three Cochrane corpora, again until convergence. The corresponding systems are post-fixed with `*-ft`.

2.2.2 Pre-translating terms

Medical terms, made of monolexical or polylexical units, are abound in medical terms, and getting their translation right is a very difficult task. Approaches to Biomedical MT have tried to deal this in various ways including explicitly using terminology list (Carrino et al., 2019), domain adaptation (Hira et al., 2019; Stojanovski et al., 2019) and transfer learning (Khan et al., 2018; Peng et al., 2019; Saunders et al., 2019).

We developed systems aimed at improving the translation of terms mainly following the recent proposals of (Dinu et al., 2019; Song et al., 2019). They mostly imply to pre-translate English terms into French, merely replacing the English version with a desired translation in a preprocessing step. The translation system thus inputs mixed-language sentences comprising both English and French words. In our implementation, we followed (Song et al., 2019) and did not mark the pre-translated segments in the input. The target side (French) remained unchanged. Figure 1 displays a sentence extracted from Medline 18 before and after pre-translation (in the latter, French segments are underlined).

Terms are extracted from the French-English version of the Medical Subject Headings thesaurus (MeSH), available in XML format.[13] We extracted a list of about 30K English terms and their preferred translation. This list was extended by searching our training corpus for instances where (a) a term is found in the English sentence; (b) a possible translation is found in the French sentence. Step (b)

[8] https://github.com/berkmancenter/mediacloud-sentence-splitter

[9] http://www.statmt.org/moses/

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

[11] https://Huggingface.co/transformers/model_doc/bert.html

[12] This process is not completely error prone, and yields a BLEU score of 98.2 on Medline 18 test set.

[13] http://mesh.inserm.fr/FrenchMesh/

ID	Train	Detail	ID	Medline			ID	Medline			ID	Medline		
				18	19	20		18	19	20		18	19	20
				Moses				HuggingFace						
X0	**wmt**	WMT data	**M0**	20.7	22.6	27.3	**H0**	26.8	29.6	33.7	**B0**	26.1	29.0	32.9
X1	**base**	All data	**M1**	24.7	25.9	32.6	**H1**	27.7	30.2	35.9	**B1**	28.6	31.1	37.2
X2	**base-ft**	X1 ⇒ X2	**M2**^{*2}	25.6	26.1	32.9	**H2**	28.1	30.0	35.5	**B2**	38.8	29.5	35.8
Back Translations of Monolingual data														
X3	**base+bt**	X1 + BT	**-**	-	-	-	**H3**	27.9	30.8	36.7	**B3**	28.0	31.0	36.3
X4	**base+bt-ft**	X3 ⇒ X4	**-**	-	-	-	**H4**^{*1}	28.7	30.7	37.0	**B4**	31.6	30.8	36.2
Using Pre-translated terms														
X5	**base+bt-pt**	X3 ⇒ X5	**-**	-	-	-	**H5**	27.5	30.0	35.9	**B5**	29.0	30.2	36.3
X6	**base+bt-pt-ft**	X5 ⇒ X6	**-**	-	-	-	**H6**	33.0	27.0	32.5	**B6**	36.0	28.8	35.2
Using IR retrieved corpus														
X7	**base+bt+IR**	X3 + IR	**-**	-	-	-	**H7**	28.8	31.4	37.2	**B7**	28.8	31.2	36.5
X8	**base+bt+IR-ft**	X7 ⇒ X8	**-**	-	-	-	**H8**	29.4	31.0	37.3	**B8**^{*3}	31.7	30.6	36.5

Table 2: BLEU scores for the various biomedical systems on Medline 18, 19 and 20 test sets. Superscripts ^{*n} denote the runs submitted: H4, M2, B8.

relies on a much larger list of about 800K possible associations, also extracted from the MeSH. The final term list contains about 40K entries.

Training was performed in two steps: starting with our best system (M3), we resume training with partially pre-translated sentences, using only the following corpora: Cochrane, Medline, Taus and a large portion of Scielo (for a grand total of 2M sentence pairs). This process is performed until convergence. The same fine-tuning process as described above is optionally performed.

In testing, we replace any matching English term with its translation subject to length constraints to avoid irrelevant, ambiguous or accidental matches. We only substitute terms of (source+target) length greater or equal to 7 characters, yielding the pre-translation of 462 and 795 terms respectively in the Medline 18 and Medline 19 test sets. Cases where one term has several translations are disambiguated based on frequency of occurrences in training. These systems appear in the last two rows of Table 2 with the postfix `*-pt`.

2.3 Translation framework

We mostly used two architectures to build our systems: basic Transformer models (Vaswani et al., 2017) as well as BERT-fused transformer models (Zhu et al., 2020). All systems use Facebook's seq-2-seq library fairseq (Ott et al., 2019) with parameters settings borrowed from `transformer_iwslt_de_en`.[14] We used memory efficient FP16 optimizer. The ReLU activation function was used in all 6 encoder and 6 decoder layers, 1024 hidden layer size and batch size of 4K. Training was optimized using Adam and a learning rate of 0.0005 was fixed for all experiments.

For the BERT-based models, we relied on BERT-NMT.[15] This allowed us to build the BERT-fused models using the same architecture and parameters as the baseline transformer models and to establish fair comparisons. In BERT-fused NMT model, the contextual representations are first computed by the BERT model for each token (in the source and target), these are then combined at each encoder and decoder layer using the attention mechanism. Full details are in Zhu et al. (2020).

Given the size of our training data, the "lazy" output dataset implementation was used to enable data loading in the RAM. Systems were trained until convergence based on the BLEU score on the development sets. Evaluation is performed using sacrebleu (Post, 2018). Scores are chosen based on the best score on the development set (Khres+Edp+Scielo) and the corresponding scores for that checkpoint are reported on Medline 18 and

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

[15] https://github.com/bert-nmt/bert-nmt

Medline 19 test sets. For systems using terminology pre-translation, Khresmoi and Edp were used as development sets.

2.4 Results

Results are in Table 2, where we report BLEU scores for the three tracks explored in this work. $M*$ denotes the Moses tokenization pipeline, $H*$ represents the HuggingFace pipeline and $B*$ denotes the BERT models with HuggingFace tokenization. We computed the scores on Medline 18, Medline 19 and Medline 20 test sets,[16] based on the best checkpoint on our development corpus. Base systems are given on the left, ($\Rightarrow$) identifies the derived (fine-tuned) systems.

We first built baseline systems for the three tracks. `X0` denotes the systems built using only the data provided by the organizers. `X1` are our baseline systems built using all our parallel corpora. We see a unanimous improvement in all tracks ranging from 0.6 to 5.3 BLEU points, which is obtained by adding around $1M$ sentences of additional Cochrane and Taus corpora to the already available $2.9M$ sentences from WMT20. This hints at the relevance of the additional in-domain parallel corpora used.

These baselines `X1` are then further fine-tuned with Cochrane and Medline abstracts as discussed in section 2.2.1, these are shown post-fixed with `*-ft`. All the systems show an improvement in the Moses track. Similarly, we see gain for all tracks for Medline 18 with the highest improvement on BERT-fused systems. For Medline 19 and 20, fine-tuning resulted in a small drop in performance across the board (except than Moses track), for reasons that remain to be analyzed.

Comparing M1-M2 with H1-H2, we see that the Moses pre-processing, which is simpler that HuggingFace's and relies on domain-adapted BPE units is slightly better than the alternative. As using HuggingFace's tools was a way to also experiment with BERT and other extensions, it was nonetheless used for the other systems.

Having established the adequacy of the supplementary parallel corpora, we built systems with back-translated monolingual corpora (section 2.1.1). These appear as X3 and X4 in Table 2. These back-translations were somewhat helpful, not to the extent that we were expecting them to be. Comparing with our baseline X1 systems, we see a small gain of (0.2,0.6,0.8) for our transformer models using HuggingFace tokenization (H1 vs. H3) but no gain for the BERT track (B1 vs. B3). We can speculate about various reasons for this behaviour: (a) genre mismatch with the test set: even though the monolingual corpora also contain scientific texts in biomedical domain, the use of full documents might yield subtle differences in style and term used with what is observed in abstracts, which are more rigidly structured; (b) the use of a comparatively small amount of back-translations as compared to the baseline corpora; (c) the quality of back translations.

Our experiments with pre-translated terms resulted in a small drop of the BLEU scores for the corresponding systems (X5, X6). Our initial analysis of term use[17] in the references and in the system outputs helps understand why this is the case. As it turns out, references translations contain a smaller proportion of *licensed terms* than our baseline translations (55.6% for the reference, 61.1% and 61.6% for respectively X3 and X4), which in turn contain less terms than our term-sensitive systems (H5 and H6, for which these numbers are respectively 68.9 and 64.2). Another way to look at this is to realize that only 58.6% of our pre-translations were actually in the reference. All in all, using more translations from the MeSH makes our output less similar to the reference than the baselines, and contributes to degrade the BLEU score. It is however reinsuring to see that pre-translating terms actually increases the number of terms in the output – in fact, for H5 and H6 we find that respectively 84.2% and 81.9% of these pre-translations are actually copied in the target, even though there was no indication of these French inserts in the mixed-language input. We can also note that the majority of the pre-translated terms were frequent Biomedical terms (such as "patients", "health", etc) that were also correctly translated by the baseline systems. Evaluating these outputs with more useful metrics than BLEU still needs to be performed.

Adding the IR retrieved sentences finally brought us nearly one extra BLEU point on all test sets for the HuggingFace systems, but not much improvement for the BERT-fused system.

[16] Again with our own sentence alignment.

[17] Based on the proportion of source word in our term list that are actually translated with a translation that exists in the Mesh. These proportions are computed on an aggregate of the Medline testsets for 2018, 2019 and 2020, only counting terms with source+target length greater than 7.

Domain	Corpus	sents.	words (en)	words (de)
web	Paracrawl	50,875	978	919
economy	Tilde EESC	2,858	61	58
news	Commoncrawl	2,399	51	47
	Tilde rapid	940	20	19
	News commentary	361	8	8
tourism	Tilde tourism	7	0.1	0.1
gov	Epps	1,828	45	42
medical	Tilde EMEA	347	5	5
banking	Tilde ECB	4	0.085	0.074
wiki	Wikipedia Matrix	5,473	91	88

Table 3: Data used in the Robustness task: number of parallel lines ($\times 10^3$), number of tokens ($\times 10^6$)

2.5 Conclusion

In conclusion, our participation to this year's WMT biomedical task has enabled us to develop basic tools and pipelines for a variety of architectures and to start exploring domain-adapted extensions of a baseline Transformer architecture, using complementary resources, such as supplementary corpora, pre-trained embeddings and terminological resources. If all these extensions were not equally useful, we still were able to develop strong systems for this task that provide us with a solid starting point for further developments of domain-adapted NMT systems.

3 Robustness: translating English challenge test sets into German

3.1 Data sources

Our sole data sources are the parallel corpora distributed by the organizers for the News task, which we significantly down-sampled in order to reduce the overall computational training cost. Monolingual data sources were not considered. These parallel corpora were then grouped into 8 broad domains. Statistics for each corpus / domain are in Table 3.

Our development set is composed of a varied set of common benchmarks, aimed to represent a wide diversity of genres and domains.

3.2 Pre-processing

The first step of pre-processing consists of cleaning the parallel corpora using the following rules: (a) discard sentences based on length (with a maximum length of 99 words), and on the source/target length ratio (in the interval $[2/3; 3/2]$); (b) discard instances of non-English and non-German sentences, using the langid toolkit;[18] (c) remove duplicates sentence pairs. After cleaning, the parallel corpus used in training contains 50,875,449 sentences pairs.

The next step is to lowercase and to tokenize the text into words and subword units. We use the Tokenizer library from OpenNMT.[19] We first lowercased every word, adding a special marker at the beginning of capitalized words, and likewise for uppercased words and segments. For instance, this procedure replaces "It" with "**U** it", and "NOVEMBER RAIN" with "**BU** november **EU BU** rain **EU**". These markers are preserved during the BPE tokenization. We learned a joint BPE vocabulary for both languages using 32K merge operations.

3.3 Training a robust multi-domain system

Our approach to robustness aims at building a system that (a) could fare well for test sets that would be similar to the training domain; (b) could also accommodate data from new, unseen, domains; (c) would be easy to adapt to a new domain (for the few-shot condition); (d) could be robust to spelling noise in the test. Requirements (a)-(c) lead us to implement an extension of the baseline Transformer architecture with residual adapters (more on this in section 3.3.2); to meet requirement (d), we implemented a data augmentation technique described in Section 3.3.3.

3.3.1 Baseline

The baseline system relies on the Transformer Large architecture from (Vaswani et al., 2017). We set the embeddings size and the hidden layers size to 1024. Transformers use multi-head attention with 16 heads in each of the 6+6 layers; the inner feedforward layer contains 4096 cells. Training uses a batch size of 12288 tokens; optimization uses Adam with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$ and Noam decay ($warmup\_steps = 4000$) and a dropout rate of 0.1 for all layers.

3.3.2 Residual adapters

Our main source of inspiration is the work of Bapna and Firat (2019), who initially introduced the use of residual adapter modules for domain adaption. In a nutshell, this proposal adds an additional, domain-specific layer on top of every layer of the encoder and the decoder. It thus provides us with

[18] https://github.com/saffsd/langid.py
[19] https://github.com/OpenNMT/Tokenizer

a lightweight, computationally efficient alternative to domain adaptation with full fine-tuning, which implies to update all the system parameters. We generalize this approach by training (or rather fine-tuning) a distinct residual adapter for each of the 8 train domains, while freezing the parameters of the baseline (generic) system. These adapter modules are made of 2-layer perceptrons, with an inner ReLU activation function operating on normalized entries of dimension 2048.

Any test sentence from a known domain would then use the corresponding adapter; for test sentences from new domains two options are possible: use only the generic system (without adapter), or use the adapter for the more similar domain. This methodology was chosen in the view of the few-shot task, where a new adapter could easily be learned for a new domain, even with a very small amount of data.

We evaluate the effectiveness of the residual adapters architecture using a varied set of internal test sets. Table 4 reports the BLEU scores of the baseline, generic model, prior to adaptation, as well as the adapted system. As expected, performance are overall better when selecting the appropriate domain for each test set.

We applied this idea to improve the ability our generic model to handle noisy data. Recall that most of the training data (with the exception of the web domain) comes from "clean" sources. To this end, we generated artificial training data for an additional "noise" domain, by automatically altering the source side of randomly selected training data. The noise generation procedure is described below. By doing this way, we expect the model to take advantage of the residual layer when input with noisy data that is similar to our artificial noisy domain, while keeping (a) its good performance on the other known domains, (b) a reasonable behaviour on any other clean data (using the generic baseline model without adapter).

3.3.3 Artificial noise generation

In order to account for possible user generated content (UGC) at test time, we explored the possibility of learning typical UGC noise at the character-level. To this end, we used an automatically scrapped Wikipedia correction corpus (Grundkiewicz and Junczys-Dowmunt, 2014), which has been filtered to keep only word replacements with, at most, a character edit distance of 30% of the word length. In the end, we kept a total of roughly 17.8M pairs of errors and editions. We then trained a character-level Transformer with the same architecture as our base translation model, which had a perfect-match error rate of 22% on the test data partition. Finally, we augmented the original training data by sampling random original words according to a uniform probability distribution and replacing them with the prediction of our character-based UGC noise generator, resulting in the same number of sentences in the original corpora. We have set a 7% probability of replacement, that has been estimated by the percentage of Out-of-Vocabulary words in a real-world UGC corpus. This heuristic later seemed, as discussed in Section 3.4, to overestimate the quantity of noise to be added and, in retrospective, we should have used other metrics to estimate the noise level, such as the n-gram Kullback-Leibler divergence, as discussed in (Alonso et al., 2016; Rosales Núñez et al., 2019). Table 5 displays some examples of noise entries produced by our character-based generator. Regarding these, although typographical errors prevail, due to the nature of automatic filtering of the Wikipedia editions, some learned replacements operations can change the semantics and syntax of the sentence, e.g. (using $\rightarrow$ use), (for $\rightarrow$ in) or (may $\rightarrow$ can); thus introducing unexpected confusion in the training data.

3.4 Results

We report the BLEU scores of our various systems in Table 6. Our submission to the zero-shot evaluation was `FT-Adapt-Noise`, which we found was sub-optimal afterwards. However, interestingly, the residual adapter mechanism proved to substantially outperform the classical fine-tuning of the whole model (i.e. `FT-Full-Noise`). Finally, the residual adapter fine-tuned using the ParaCrawl corpus (`FT-Adapt-Web`) had the best performance on the test set, probably due to the higher similarity of this corpus to the target test. In addition, we noted that the baseline and `FT-Adapt-Noise` output a considerable number of English phrases, leaving most of the source sentence unchanged, whereas the `FT-Adapt-Web` reduced the number of sentences that presented this issue.

In order to assess how much the 172 sentences that were left completely untranslated impact the performance of the `FT-Adapt-Noise` model, we replaced them with the output of the

Test set Domain	IT tech	Khresmoi medical	NT17	NT18 news	NT19	EPPS gov	EESC eco	RAPID news	Tourism tourism	Wiki wiki	ECB bank
Baseline	36.27	29.78	26.24	41.27	37.24	29.31	30.48	31.93	17.64	14.92	38.11
FT-Adapt domain	-	29.46	26.48	41.43	37.24	29.65	30.45	32.43	19.21	-	48.99

Table 4: BLEU scores on various test sets using our baseline and adapted NMT systems for each domain. *NT stands for NewsTest*

original	the	combination	may	concerning	using	no	common	developing	for	status	also
noisy	this	combonation	can	concering	use	not	comon	developping	in	staus	aslo

Table 5: Examples of clean and artificially noisy word inputs

`baseline` and observed a performance increase to 31.3 BLEU. This suggests that our data augmentation technique introduced confusion to the base model after fine-tuning and the resulting translation system was less adapted to the zero-shot test set.

	robustness-set1	#EN Sents.
Baseline	31.6	120
FT-Adapt-Noise	30.2	172
FT-Full-Noise	24.6	256
FT-Adapt-Web	34.2	34
FT-Full-Web	33.8	49

Table 6: BLEU scores for the EN-De models developed for the Robustness track. We also report, for each system, the number of sentences that were left unchanged.

The design and organization of the few-shot part of the evaluation was not fully satisfactory: while we did train an adapter module using the new data seemingly corresponding to a novel domain, it seems that the corresponding test set was never released and we could not fully evaluate our approach. Working on this task was nonetheless very instructive, and helped us better understand the strength and pitfall of the residual adapter architecture when applied to a very large scale task and in the face of unbalanced, heterogeneous, training data.

4 Conclusions

In this paper, we have described the development undertaken for this year's participation to WMT shared tasks. Taking part to the Biomedical track as allowed us to collect and prepare useful resources (monolingual and bilingual corpora, term lists) for this domain, and to explore several pipelines and translation architectures. The general results are overall satisfactory, even though a deeper analysis of the MT is still needed to strengthen our conclusions. They will also help us prepare for next year tasks, where we expect to work on more language pairs. Our experiment for the Robustness track were less successful: we were not really prepared for the general tone and style that was observed in the zero-shot test set; we also did not understand the general orientation taken for the few-shot adaptation, as it seemed to us that the adaptation data was not really relevant for the only test set that was ever released.

Acknowledgments

This work is (partly) based on computations performed on the Saclay-IA and on the Jean ZAY computing platforms. The authors wish to thank Pierre Zweigenbaum for his help finding French corpora in the biomedical domain and Hicham El-Boukkouri for providing guidance setting up BERT-based systems. The second author wishes to acknowledge the help and guidance of Djamé Seddah and Guillaume Wisniewski; his work is funded by the French Research Agency via the ANR project ParSiTi (ANR-16-CE33- 0021).

References

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.

Héctor Martínez Alonso, Djamé Seddah, and Benoît Sagot. 2016. From noisy questions to minecraft texts: Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy*

User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016, pages 13–23. The COLING 2016 Organizing Committee.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Casimiro Pio Carrino, Bardia Rafieian, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. Terminology-aware segmentation and domain feature for the WMT19 biomedical translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 151–155, Florence, Italy. Association for Computational Linguistics.

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.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing – Lecture Notes in Computer Science*, volume 8686, pages 478–490. Springer.

Noor-e Hira, Sadaf Abdul Rauf, Kiran Kiani, Ammara Zafar, and Raheel Nawaz. 2019. Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 156–163, Florence, Italy. 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.

Abdul Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. 2018. Hunter NMT system for WMT18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 655–661, Belgium, Brussels. Association for Computational Linguistics.

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.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proc. AMTA'02*, Lecture Notes in Computer Science 2499, pages 135–144, Tiburon, CA, USA. Springer Verlag.

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).

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.

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.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT performance for translating noisy user-generated content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2019. UCAM biomedical translation at WMT19: Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 169–174, Florence, Italy. 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.

Felipe Soares and Martin Krallinger. 2019. BSC participation in the WMT translation of biomedical abstracts. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 175–178, Florence, Italy. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC'12, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In 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.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating BERT into Neural Machine Translation. In *Proceedings of the International Conference on Learning Representations*, ICLR.