The SPECTRANS System Description for the WMT22 Biomedical Task

Nicolas Ballier[1], Jean-Baptiste Yunès[2], Guillaume Wisniewski[3],
Lichao Zhu[3], Maria Zimina-Poirot[1]

[1]CLILLAC-ARP, [2]IRIF, [3]LLF
Université Paris Cité, F-75013 Paris, France
{nicolas.ballier, guillaume.wisniewski, jean-baptiste.yunes,
lichao.zhu, maria.zimina-poirot}@u-paris.fr

Abstract

This paper describes the SPECTRANS submission for the WMT 2022 biomedical shared task. We present the results of our experiments using the training corpora and the JoeyNMT (Kreutzer et al., 2019) and SYSTRAN Pure Neural Server/ Advanced Model Studio toolkits for the language directions English to French and French to English. We compare the predictions of the different toolkits. We also use JoeyNMT to fine-tune the model with a selection of texts from WMT, Khresmoi and UFAL data sets. We report our results and assess the respective merits of the different translated texts.

1 Introduction

For this WMT22 Biomedical workshop, we focused on the selection of texts used for fine-tuning. We selected what we believe to be the two best models we produced for the EN-FR track with two different neural toolkits but we mostly took the opportunity to discuss the translated texts. The rest of the paper is organised as follows: Section 2 summarises our approaches to the task, Section 3 details the training data of our experiments, Section 4 presents the results. Section 5 discusses them.

2 Our Approaches to the Task

This section presents our various strategies for this task and our four submissions. We compared the predictions of two toolkits but our comparison is very partial as the training data differs. We trained several systems with JoeyNMT (Kreutzer et al., 2019) training and fine-tuning with UFAL, WMT and Khresmoi data. We used the SYSTRAN Pure Neural® Server generic system and tried to fine-tune with specialised terminology. We used SYSTRAN Advanced Model Studio® to fine-tune a generic model with in-house data based on 2,700 aligned segments collected during the translation of the French federation for diabetes.[1] Table 1 summarises our submissions.

With JoeyNMT, we selected the training data, comparing the performance with and without the added data and applied fine-tuning to the model based on UFAL medical corpora The following section details the model selection and fine-tuning.

3 Data and Tools Used

In this section, we present different approaches that we adopted to train baseline models and proceed to fine-tuning. We have built two baseline models : one trained with generic data set fine-tuned with in-domain data, and the other trained directly with in-domain data, in order to compare their performances and to better understand functioning of in-domain NMT training.

3.1 Data for baseline models training

We used two baseline models : the first one is built based on our model submitted for WMT 2021. It took the Europarl 7 parallel corpus as data set trained with 341,554 sentences in two directions (EN⇔FR)(Ballier et al., 2021); the second one has been built by using bilingual (EN-FR) in-domain parallel corpora data set UFAL provided by WMT 2022 (with 2,693,509 sentences). The corpora have been normalized and sentences longer that 50 words have been removed. Thus we have retained 2,159,307 sentences. These sentences are split in the ratio of 6-2-2 : 60% for training, 20% for development and the last 20% for evaluation. Two tokenizations are applied to all the data sets : standard tokenization (`Spacy`) segments data into words and BPE tokenization into sub-words with `SentencePiece` (Kudo, 2018).

[1] https://www.federationdesdiabetiques.org. Diabetes terminology proved to be not so useful for the actual test set.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 895–900
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

run	BLEU (into English)	BLEU (from English)	toolkit	training data
run1	0.2581	0.2068	JoeyNMT	baseline with UFAL
run2	0.4010	0.31636	Pure Neural Server	general training data
run3	0.2587	0.0732	ModelStudio Light	fine-tuning with in-house data
run4	0.0969	0.2034	JoeyNMT	UFAL fine-tuned

Table 1: Summary of our official submissions

3.2 Data for fine-tuning

We used two data sets to fine-tune the generic baseline model. For the first data set, we have compiled the WMT Medline parallel corpus since 2016 [2] as well as Khresmoi dev and test data (EN-FR) [3]. The whole data set contains 109,912 sentences. For the second one, we used the normalized and sub-tokenized UFAL data set mentioned above.

4 Experiments and Results

In our experiments, we aimed to compare the different JoeyNMT models (baseline and fine-tuning) that we have trained with SYSTRAN model. JoeyNMT, which is based on TRANSFORMER (Vaswani et al., 2017), requires lighter implementations than OpenNMT (Klein et al., 2017).

4.1 Baseline with JoeyNMT

We have trained a baseline model with in-domain data set UFAL. For FR→EN model, the best checkpoint is recorded at step 60,000 with a BLEU score of 61.01 (PPL: 1.53); as for EN→FR model, the best checkpoint is recorded at step 40,000 with a BLEU score of 59.23 (PPL: 1.45, see Figure 1) [4].

4.2 Fine-tuning with JoeyNMT

The generic baseline model was fine-tuned with the following parameters: vocabulary size: 32,000, maximum sentence length: 50, maximum output length: 100, training initializer: XAVIER, number of layers: 6, number of heads: 8 normalization: tokens, encoder embedding dimension: 512, decoder embedding dimension: 512, hidden size: 512. It was fine-tuned with two data sets. The first one with Medline-Khresmoi data set got the best BLEU score from French to English 54.8, 38.4 as from English to French (see Figure 2).

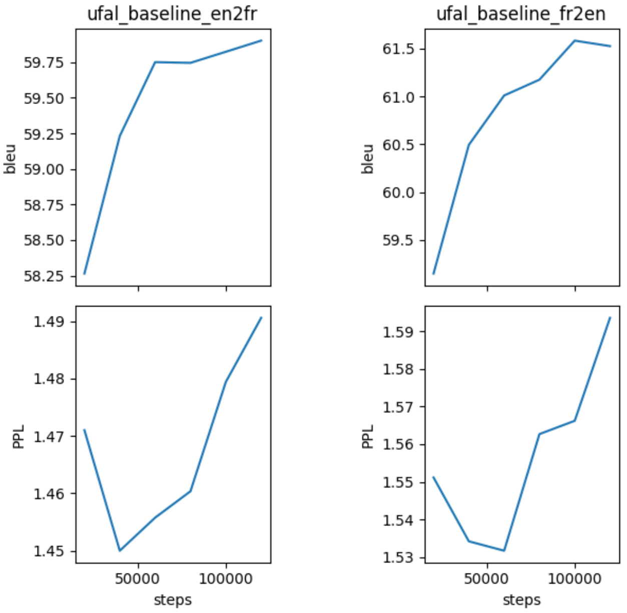


Figure 1: Baseline trained with UFAL data set FR⇔EN

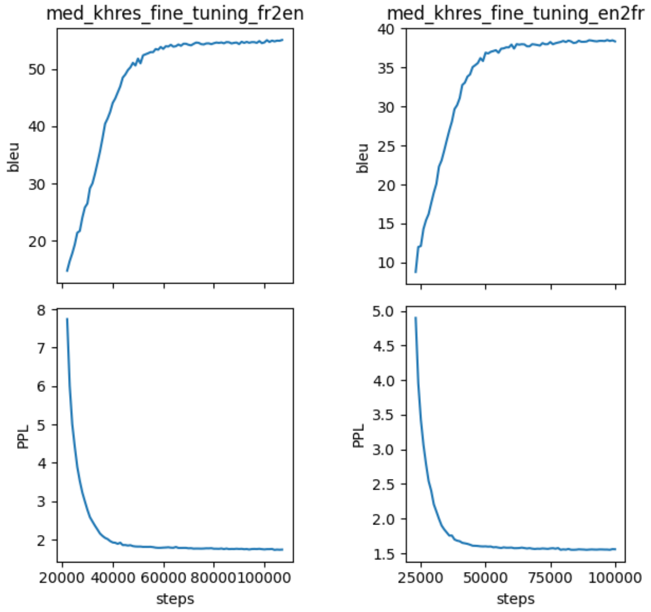


Figure 2: Fine-tuning with Medline and Khresmoi data set FR⇔EN

[2] https://github.com/biomedical-translation-corpora/corpora

[3] http://hdl.handle.net/11234/1-2122

[4] We noticed that the validation processes were extremely long. Every validation after 20,000 steps took about 28 hours.

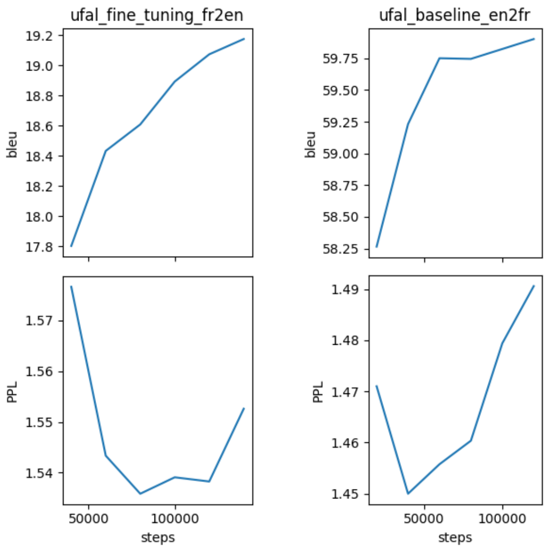


Figure 3: Fine-tuning with UFAL data set FR⇔EN

With the same parameters, the model fine-tuned with UFAL data set had, surprisingly, relatively low scores : we obtained a BLEU score of 18.60 for French→English model and 21.13 as for English→French model (see Figure 3).

4.3 Training and Fine-tuning with Systran Model Studio

SYSTRAN Pure Neural® Server is a multilingual translation platform that offers website translation and localisation features. [5] The server uses Pure Neural® Machine Translation (PNMT®), a commercial engine based on AI and deep learning, launched in 2016. This technology enables neural engines to learn language rules from a given translated text and to produce a translation achieving the current state of the art. An open source neural machine translation system OpenNMT developed by the Harvard NLP group and Systran is available online: http://opennmt.net.

For our work, we used SYSTRAN Pure Neural® Server installed on PAPTAN [6].

We used *characteristic elements* computation (Lebart et al., 1997) implemented in *iTrameur*[7] to compare the results of run2 (generated by SYSTRAN Pure Neural® Server) and run1 (generated by JoeyNMT), using characteristic elements computation (Lebart et al., 1997). In this paper, we discuss the results of FR→EN translation (Table 2). As one can see in Table 2, in the SYSTRAN translation, a sentence always starts with capitalization (“The”, “This”, “In”). Capital letters are also used for acronyms and abbreviations (“BMI”, “VCE”). This can be explained by the default detokenization function of JoeyNMT in detokenizing translation in sub-tokenized form.

Unit	Fq part	Fq total	IndSP
The	108	108	+33
This	36	36	+12
In	39	39	+12
vaping	28	28	+9
We	23	23	+8
It	19	19	+7
These	18	18	+6
They	16	16	+6
Finally	14	14	+5
must	14	14	+5
Management	9	9	+4
advances	10	10	+4
BMI	11	11	+4
Cancer	10	10	+4
liaison	10	10	+4
However	9	9	+4
VCE	10	10	+4
we	22	71	-4
search	0	10	-4
gc	0	10	-4
bmi	0	13	-5
the	659	1469	-7

Table 2: Characteristic elements of Systran translation (run2) and JoeyNMT translation (run1)

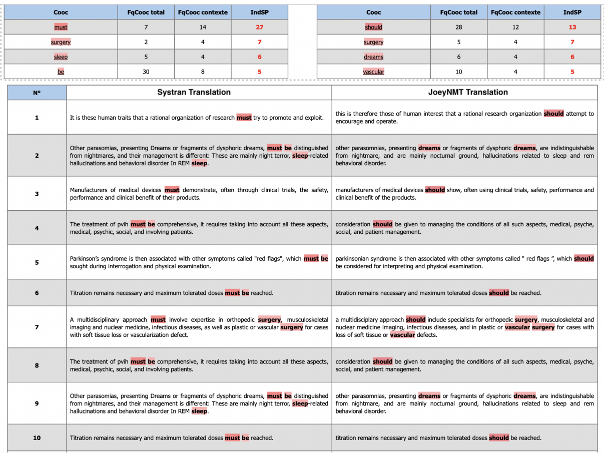
The modal verb “must” is overused in the SYSTRAN translation (IndSP = +5) and is never used in the JoeyNMT translation, which tends to prefer the use of the modal verb “should” (Figure 4). The absence of "must" produced by the JoeyNMT system might be due to the large difference of frequencies of both words in training data : 18,462 occurrences of "should" and 4,061 occurrences of "must". The preponderance of "should" in the training corpus has seemingly induced the system to systematically produce the word whenever the system needs to produce a modal verb before a base verb.

We also note that JoeyNMT translation under-

[5] The official product website is available at https://www.systransoft.com/translation-products/systran-pure-neural-server/.

[6] Plateforme pour l’apprentissage profond pour la traduction automatique neuronale, in English: Deep Learning for Machine Translation at Université de Paris-Cité). See the description of the platform on the project website: https://u-paris.fr/plateforme-paptan.

[7] https://itrameur.clillac-arp.univ-paris-diderot.fr



Cooc	FqCooc total	FqCooc contexte	IndSP
must	7	14	27
surgery	2	4	7
sleep	5	4	6
be	30	8	5

Cooc	FqCooc total	FqCooc contexte	IndSP
should	28	12	13
surgery	5	4	7
dreams	6	4	6
vascular	10	4	5

N°	Systran Translation	JoeyNMT Translation
1	It is these human traits that a rational organization of research **must** try to promote and exploit.	this is therefore those of human interest that a rational research organization **should** attempt to encourage and operate.
2	Other parasomias, presenting Dreams or fragments of dysphoric dreams, **must be** distinguished from nightmares, and their management is different: These are mainly night terror, **sleep**-related hallucinations and behavioral disorder In REM **sleep**.	other parasomnias, presenting **dreams** or fragments of dysphoric **dreams**, are indistinguishable from nightmare, and are mainly nocturnal ground, hallucinations related to sleep and rem behavioral disorder.
3	Manufacturers of medical devices **must** demonstrate, often through clinical trials, the safety, performance and clinical benefit of their products.	manufacturers of medical devices **should** show, often using clinical trials, safety, performance and clinical benefit of the products.
4	The treatment of pvih **must be** comprehensive, it requires taking into account all these aspects, medical, psychic, social, and involving patients.	consideration **should** be given to managing the conditions of all such aspects, medical, psyche, social, and patient management.
5	Parkinson's syndrome is then associated with other symptoms called "red flags", which **must be** sought during interrogation and physical examination.	parkinsonian syndrome is then associated with other symptoms called " red flags ", which **should** be considered for interpreting and physical examination.
6	Titration remains necessary and maximum tolerated doses **must be** reached.	titration remains necessary and maximum tolerated doses **should** be reached.
7	A multidisciplinary approach **must** involve expertise in orthopedic **surgery**, musculoskeletal imaging and nuclear medicine, infectious diseases, as well as plastic or vascular **surgery** for cases with soft tissue loss or vascularization defect.	a multidisciplary approach **should** include specialists for orthopedic **surgery**, musculoskeletal and nuclear medicine imaging, infectious diseases, and in plastic or **vascular surgery** for cases with loss of soft tissue or **vascular** defects.
8	The treatment of pvih **must be** comprehensive, it requires taking into account all these aspects, medical, psychic, social, and involving patients.	consideration **should** be given to managing the conditions of all such aspects, medical, psyche, social, and patient management.
9	Other parasomias, presenting Dreams or fragments of dysphoric dreams, **must be** distinguished from nightmares, and their management is different: These are mainly night terror, **sleep**-related hallucinations and behavioral disorder In REM **sleep**.	other parasomnias, presenting **dreams** or fragments of dysphoric **dreams**, are indistinguishable from nightmare, and are mainly nocturnal ground, hallucinations related to sleep and rem behavioral disorder.
10	Titration remains necessary and maximum tolerated doses **must be** reached.	titration remains necessary and maximum tolerated doses **should** be reached.

Figure 4: Comparison occurrences of "must" and "should" in SYSTRAN and JoeyNMT translations

SYSTRAN translation	JoeyNMT translation
We take stock of knowledge about this addiction and its management.	knowledge about this dependency and the management thereof is a pending state.

Table 3: "we" in SYSTRAN and JoeyNMT translations

uses “we” (IndSP = -4). This finding is interesting because it makes sometimes possible to identify substantial differences between both translations in Table 3.

These results show how training data affects translation results. To our knowledge, SYSTRAN NMT relies upon a broad selection of general texts that do not belong to any single text type, subject field, or register (many of them are translated texts from the web available on https://opus.nlpl.eu). The WMT corpus consists of randomly selected sentences from abstracts and main texts of scientific articles published in medical journals. The articles follow the so-called introduction, methods, results and discussion structure (IMRAD) (Heßler et al., 2020). The selection is not necessarily balanced in terms of represented discourse functions. Thus, we noticed the overuse of “should be” that definitely constrained our translation output (see Figure 4 “should be given”, “should be reached”, “should be considered”, etc.).

5 Discussion

5.1 Degrees of Specialisation

If the Biomedical terminology was indeed present in the testing set (eg "hypertension artérielle pulmonaire","nutriments", "supplémentation en vitamine D"), some sentences were not particularly specialised. For instance, "Le but de cet article est de les résumer de manière relativement exhaustive." is representative of Scientific French for specific purposes but not really of biomedical specialised language. The same holds for the test set from English into French. In view of these observations, it is easy to understand why models trained on more generic data perform so well in this task.

5.2 The performance of gigamodels

We have not submitted translations produced on `mBART-50` (Tang et al., 2021), but we compared the translations of our best system (PNS for Pure Neural Server) with those of mBART. [8]. The translation based on mBART produces fluent grammatical sentences but seems to be less specific in the terminology. For instance "vapotage" (*vaping testing*) was translated as *poultry testing* and instead of *vaping frequency* the system produced *pooping frequency*. The terminology is not always consistent or accurate : *hyperthyroïdie frustre* was translated as *rough* (SYSTRAN) or *fruity* (MBART). Oddly enough, with mBART, percentages were literally translated as "per cent" instead of the % symbol.

Figure 5 plots the vocabulary growth curves (VGCs) of the two translated texts. The `y` axis corresponds to the number of new types and the `x` axis corresponds to the number of tokens in the translated texts. As can be seen, the two systems have remarkably similar patterns of VGCs, with SYSTRAN PNS slightly above MBART, in spite of the variants we noticed. For the French translation of "keloids", mBART varies between "céloïdes" and "keloïdes", whereas SYSTRAN PNS only produces "chéloïdes".

Measuring specificity indices (Lebart et al., 1997) allowed us to spot differences in the translation. One of the most striking ones was the choice of feminine determiner *la* for *la COVID* in the PNS translations, as evidenced by the specificity of *la COVID* in the two translations (Figure 6). A somewhat belated and debated ruling of the Académie

[8] We used `mBART1` through the `HuggingFace` API (Wolf et al., 2020).`https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt`

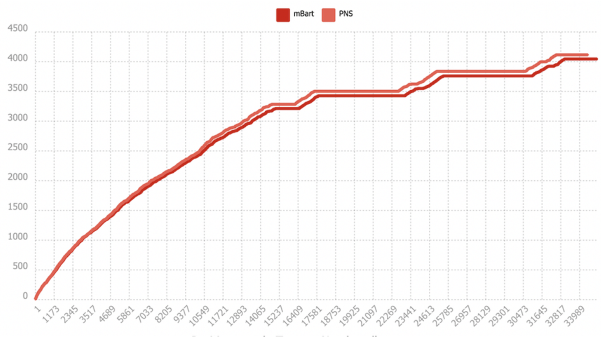


Figure 5: Comparison of Vocabulary Growth Curves in SYSTRAN PNS and mBART translations

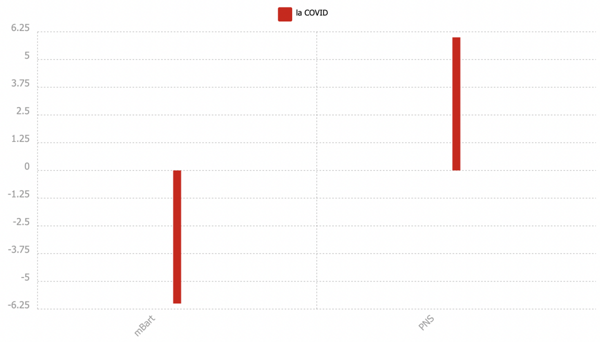


Figure 6: Comparison of Specificity Vocabulary Growth Curves in Systran PNS and mBART translations

française endorsed and imposed "*la*" for the gender of COVID in French. This benign detail probably can be used as a chronological landmark for the training data collection of the two systems: it seems that PNS was trained with more recent French texts. It may also be the case that SYSTRAN has used rule-based normalisation to regularise the output for *la COVID*.

6 Conclusion

This paper presents the SPECTRANS system description for the WMT 2022 biomedical Shared Task. We participated in the English-to-French and French-to-English tasks. We only used the data provided by the organisers but also analysed the translations produced with mBART. We obviously concur with previous research that training data is key. For the MT system, we applied a variety of strategies, toolkit comparison and fine-tuning to compare outcomes of different NMT systems in biomedical translation.

Our contribution mostly lies in the textometric analysis of the output. This allowed us to raise the issue of the role of the variability observed for the gender of COVID in French or for technical terms like "keloids".

Acknowledgements

The SPECTRANS project is funded under the 2020 émergence research project and this publication has emanated from research supported in part by a 2021 research equipment grant from the Scientific Platforms and Equipment Committee (PAPTAN project), both under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris).

We gratefully acknowledge support from the CNRS/TGIR HUMA-NUM and IN2P3 Computing Center (Lyon - France) for providing some of the computing and data-processing resources needed for this work.

References

Nicolas Ballier, Dahn Cho, Bilal Faye, Zong-You Ke, Hanna Martikainen, Mojca Pecman, Jean-Baptiste Yunès, Guillaume Wisniewski, Lichao Zhu, and Maria Zimina-Poirot. 2021. The SPECTRANS System Description for the WMT21 Terminology Task. In *EMNLP 2021 SIXTH CONFERENCE ON MACHINE TRANSLATION (WMT21)*, Proceedings of the Sixth Conference on Machine Translation, pages 815–820, Punta Cana, Dominican Republic. ACL.

Nicole Heßler, Miriam Rottmann, and Andreas Ziegler. 2020. Empirical analysis of the text structure of original research articles in medical journals. *PLOS ONE*, 15(10):1–10.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. 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): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Ludovic Lebart, André Salem, and Lisette Berry. 1997. *Exploring textual data*, volume 4. Kluwer Academic.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. 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 *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.