# The SPECTRANS System Description for the WMT21 Terminology Task

Nicolas Ballier, Dahn Cho, Bilal Faye, Zong-You Ke, Hanna Martikainen, Mojca Pecman, Jean-Baptiste Yunès, Guillaume Wisniewski, Lichao Zhu, Maria Zimina-Poirot

# The SPECTRANS System Description for the WMT21 Terminology Task

**Nicolas Ballier**[1] **Dahn Cho**[1] **Bilal Faye**[1] **Zong-You Ke**[1] **Hanna Martikainen**[2] **Mojca Pecman**[1]
**Jean-Baptiste Yunès**[3] **Guillaume Wisniewski**[4] **Lichao Zhu**[4] **Maria Zimina-Poirot**[1]
[1] CLILLAC-ARP / [3] IRIF / [4] LLF, Université de Paris, F-75013 Paris, France
[2] CLESTHIA, Université Sorbonne Nouvelle - Paris 3, F-75005 Paris, France
{nicolas.ballier, guillaume.wisniewski,jean-baptiste.yunes, lichao.zhu
maria.zimina-poirot}@u-paris.fr, mpecman@eila.univ-paris-diderot.fr
{dcho0501, biljolefa, zongyou.ke.fr}@gmail.com
hanna-julia.martikainen@sorbonne-nouvelle.fr

## Abstract

This paper discusses the WMT 2021 terminology shared task from a "meta" perspective. We present the results of our experiments using the terminology dataset and the OpenNMT (Klein et al., 2017) and JoeyNMT (Kreutzer et al., 2019) toolkits for the language direction English to French. Our experiment 1 compares the predictions of the two toolkits. Experiment 2 uses OpenNMT to fine-tune the model. We report our results for the task with the evaluation script but mostly discuss the linguistic properties of the terminology dataset provided for the task. We provide evidence of the importance of text genres across scores, having replicated the evaluation scripts.

## 1 Introduction

In our (traditional) sense, terminological databases are the collection of specialised lexical resources that are generally compiled from corpora, in collaboration with experts from the field, then analysed and structured according to the type of information recorded in term records: terms, equivalents, definitions, synonyms, contexts of use, and related terms (hyperonyms, hyponyms, meronyms, holonyms, etc.). The data thus created are empirical and provide knowledge-based representations of the domain (especially in the case of an ontological approach), underlining conceptual links between terms that can be observed (like meronomy: "X is a part of Y") and potentially represented in conceptual graphs.

For instance, the ARTES database (Pecman and Kübler, 2011), used at Université de Paris in Masters studies for teaching terminology management to translation students (Kübler et al., 2018), adopts such a comprehensive approach to terminology, with specific attention to emerging terminology and complex noun phrases (CNPs) (Kübler et al., 2021). In recent works combining studies on terminology, specialised translation and corpus linguistics, attention has been drawn to CNPs in English which have been demonstrated to cause major difficulties during translation, both human and machine (Kübler et al., 2021; Maniez, 2017). Moreover, studies have demonstrated an increase of complex compounding in specialised texts in English over the last few decades, with, for instance, an overwhelming use of patterns with adjectival and participial compound pre-modifiers (e.g. *receptor-binding activity, electron-dense aggregates*) (Mestivier-Volanschi, 2015).

For this WMT21 Terminology workshop, we focused on the linguistic properties of the terminological dataset provided. 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 addition of terminology to neural machine translation. The rest of the paper is organised as follows: section 2 summarises our approaches to the task, section 3 presents the tools we used and how we used the constrained data. Section 4 presents our experiments and the best models we used for the translation challenge. Section 5 discusses our results.

## 2 Our Approaches to the Task

This section presents our various strategies for the terminology task.

### 2.1 Toolkit Comparison

We compared the predictions of two toolkits. We trained two systems, JoeyNMT (Kreutzer et al., 2019) and OpennMT (Klein et al., 2017) with comparable parameters, using Europarl as baseline, later supplemented with the terminology resource provided for the task.

### 2.2 Model selection and fine-tuning

With OpenNMT only, we selected the training data, comparing the performance with and without the

terminological data for CommonCrawl and Europarl and applied fine-tuning to the model based on Europarl enriched with the terminological data.

### 2.3 Comparing with pre-trained models

We were curious to see how pre-trained models fared on this tasK. We produced two translations, one based on `mBART-50` (Tang et al., 2020) and the other one on the `Hugging Face` (Wolf et al., 2019) baseline. We finalised them after the evaluation deadline, so that we report our findings on the sacreBLEU score we calculated with the Systran translation used as reference. Debatable as it may sound to use an MT-generated reference translation, this enabled us to run comparisons.

### 2.4 A linguistic analysis of the terminology resource and evaluation script

We focused our analysis on the linguistic properties of the terminology provided and tested. We also tried to test other models we produced after the competition deadline, which is why we detail the evaluation script we tried to replicate and the terminology in the next section.

## 3 Data and Tools Used

This section presents the datasets used to build our system as well as our replication of the evaluation script to analyse the models we did not have the time to submit.

### 3.1 Training Data

The first challenge lies in the data selection for the training corpora among the possibilities of the challenge. We did not resort to specific texts such as the TICO-19 data (Anastasopoulos et al., 2020) but used the Europarl corpus as baseline.

### 3.2 Terminological Data

This subsection provides a linguistic qualitative approach to the provided terminology dataset.

A potential problem with the terminology dataset is variation. While some variants are probably interchangeable in most texts (e.g. 220 *hand sanitizer: gel hydroalcoolique* | *désinfectant pour les mains*), others present different degrees of specialization (e.g. 345 *multi-organ failure: défaillance multi-viscérale* | *défaillance de plusieurs organes*). For yet other variants, both forms are possible, but not within the same text for coherence (e.g. 286 *SARS-CoV-2: SRAS-CoV-2* | *SARS-CoV-2*, where the first variant is the translated acronym and stands for *syndrome respiratoire aigu sévère*).

The French-English terminological resource included 595 "terms" out of which only 181 were tested in the script so that the achievement rate as testd by the evaluation scripts only relies on 30.42% of the resource provided. Many entries in the dataset are not actually terms, but rather out-of-context strings or keyword combinations that are impossible to translate since, in translation, context truly is everything. Strings such as (154) *covid-19 WHO* and (158) *covid19 CDC* are not actual NPs and are rarely found as such, on their own, in real texts. In context, these n-grams are always followed by additional information that needs to be taken into account in their translation (e.g. *Covid-19 WHO Situation Report* or *Covid-19 CDC Info*). [1] Therefore, the proposed translations (respectively, *OMS et covid-19* and *CDC et covid19*), where the different elements are simply linked with the conjuntion *et* cannot work in context since the components of the actual NP would need to be reorganized in translation when unpacking the informational content in these CNPs. Other examples of out-of-context keyword combinations in the dataset are entries 112 *covid-19 dangerous*, 113 *covid-19 deadly*, 116 *covid19 domestic travel*, and 128 *covid19 international travel*. The role of Complex Noun Phrases seems to be underestimated in the terminology resource, as well as collocations. Nouns are more frequent than adjectives and verbs in the provided resource. 143 adjective + noun collocations are proposed (such as *deadly virus*) for 13 adjectives. Only 19 verbal collocations are proposed for eight verbs.

Beyond the immediate textual context, lack of real-world context is also a potential source for incorrect translation. For entries 245 *n95*, 246 *n95 mask*, and 247 *n95 respirator*, the proposed translations all use the N95 classification, which is the US NIOSH standard. For real texts, functionally adequate translation might require, for instance, using the equivalent European classification (*FFP2*). Dataset entry 246 presents an additional real-word related issue: *N95 respirators* should not be referred to as "masks", as their airborne-particle filtration capacity is far superior to those of surgical

[1] With hindsight, setting values at n=2 or n=3 for Window Overlap Accuracy was consistent with "truncated" sequences such as *covid-19 WHO* but *Covid-19 WHO Situation Report* and similar embedding structures would only be captured by the Window Overlap Accuracy metric when n=4 or more.

masks which serve a different purpose (reducing outward particle emission).

### 3.3 Data for Fine-Tuning

We re-trained our generic model by selecting the presumed best candidates for training sets. To specialize the model and make it more efficient, after having trained it on Europarl, we chose a method to select texts that are closer to the terminological data. Several similarity measurement methods are possible. In this case we worked with cosine similarity, which is more sensitive to the number of occurrences of terms in each corpus. After having carried out the similarity measurement of all the texts with the test data, we retained 1/4 of the files, corresponding to 22,741,561 sentences. These selected texts served as a corpus of re-training of our model for its specialization. Compared to the constrained corpora proposed for training, our optimised selection of texts based on the cosine similarity with the testing set corresponded to the following subsampling of the proposed corpora: 1 % of News Commentary v16 and 99 % of $10^9$ French-English Corpus. From a purely machine learning perspective, using testing sets to figure out training sets may sound unusual, but it should be borne in mind that we do not aim here at generalisability but at performing a specific task (translating biomedical texts).

### 3.4 Replicating the evaluation script

We did not have the time to submit our translations based on fine-tuning and pre-trained systems, so that we tried to replicate the evaluation script[2]. Our script[3] is a modification of the procedure described in (ibn Alam et al., 2021) that includes 1-TER but not COMET. It allows the calculation of the following scores: Exact-Match accuracy, Window Overlap (2), Window Overlap (3), sacreBLEU, TER and TERm. The calculation, unlike the evaluation made by the competition, is done here segment by segment and the average of the set of results makes it possible to detail scores per segments. The preprocessing is the same as on the reference script (tokenization and lemmatization) and the removal of parentheses on the corpus is necessary to run it. It is limited to 1,371 segments, for which the term to be translated was identified with certainty. As a result, one section of the testing data was not considered (the email sent to the wikipedia collaborators, referred to as "email" in our text genre analysis).

[2]https://github.com/mahfuzibnalam/terminology$_evaluation$

[3]To be found on https://github.com/nballier/SPECTRANS/tree/main/WMT21

## 4 Experiments and Results

### 4.1 Training with JoeyNMT

For comparison purposes, we used the baseline of JoeyNMT which is based on TRANSFORMER (Vaswani et al., 2017) and requires lighter implementations. It took the `Europarl 7` parallel corpus as data set, split as follows: training (341,554 sentences), dev (50,000 sentences) and test (100,000). The data set has been preprocessed with a two-level tokenization: standard tokenization (Spacy) segments data into words and BPE tokenization (*SentencePiece* (Kudo, 2018)) into sub-words. Our model was trained with the following parameters: vocabulary size: 32, 000, maximum sentence length: 50, maximum output length: 100, training optimizer: `ADAM`, normalization: tokens, training model initializer: `XAVIER`, encoder embedding dimension: 512, decoder embedding dimension: 512, hidden size: 512. The best BLEU score from English to French (Figure 1) was achieved at 32.04 at step 41 000 with a training rate of 18 seconds per 100 steps, whereas the best French to English BLEU score was 31.35. By comparing JoeyNMT translation with OpenNMT translation, we notice that JoeyNMT had poor results in translating dates, numbers, proper nouns, acronyms and symbols. Sentences which have several of those may have been translated into a string of characters of repeated sub-words. The translation submitted could not be scored but for BLEU (5.29). The result came as a surprise to us since JoeyNMT has the same model architecture as OpenNMT (Transformer). Because of these issues, we only conducted the other experiments with OpenNMT.

### 4.2 Training and Fine-tuning with OpenNMT

We used the baseline of OpenNMT-tf 2.20.1 based on `TRANSFORMER` (Vaswani et al., 2017). The parallel data Europarl v10 (Koehn et al., 2005) containing 1,911,202 aligned sentences pairs was used as a dataset, which was divided into two subsets: training set (1,906,202 sentences) and evaluation set (5,000 sentences). The dataset was preprocessed with a BPE tokenization using SentencePiece into subword units (32,000 subword units as training vo-

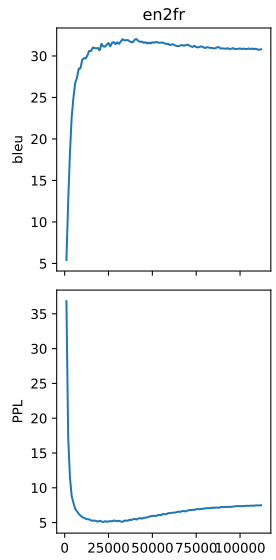


Figure 1: JoeyNMT : BLEU score and PPL score (en-fr)

cabulary). The model was trained with the following parameters: vocabulary size: 31,000, learning optimizer: LazyAdam. The best BLEU score from English to French was 43.90 after 70,000 steps with a training rate of 1.18 steps per second.

We then produced a model with Europarl adding the terminology to the training data with the same evaluation data. As a comparison, we also tried to produce a model with Common Crawl corpus [4] using the same parameters of SentencePiece and training. The dataset consists of 3,244,152 aligned sentences pairs split into training set (3,239,152 sentences) and evaluation set (5,000 sentences). This model produced the best scores in among our submissions (0.871 for Exact-Match Accuracy).

For fine-tuning, we used the Europarl model enriched with the terminological data. We were not able to use the onmt-update-vocab command, so that instead we directly replaced the dictionary file in the configuration with the dictionary based on the files described in section 3.3. Contrary to our expectations, the fine-tuning did less well for scores, according to our estimations (see Table 1 ). Not being able to update the dictionary in fine-tuning might be responsible for worsening the quality of our results.

### 4.3 Pre-trained Systems

For a point of comparison, we considered two Transformer-based models available in the Hugging Face library (Wolf et al., 2019). The first one is the standard *pipeline*[5] for English to French translation. The second one is based on the multilingual language model `mBART-50` (Tang et al., 2020), fine-tuned for multilingual machine translation as described in (Tang et al., 2020). The two models were applied on the raw sentences extracted from the SGM files of the test data. The sole preprocessing that was applied consisted in replacing XML entities by their corresponding characters and applying the tokenizer considered by the model. While the translation for the PUBMED section is satisfactory, the translation of the CMU section revealed issues in the use of subjunctive (ie segment 20). It should be noted that, according to our homemade evaluations, these models did much better for sacreBLEU scores (+3.7) and Hugging Face is slightly higher than the Corpus Crawl data trained with the terminology resources (the two models are superimposed on Figure 3.

### 4.4 Replicating the scoring system with the different translations

Because we could not submit all our translations in time, we resorted to a proxy for evaluation by adapting the available scripts to produce our own evaluation scripts. Our sacreBLEU (Post, 2018) score was based on the SYSTRAN translation used as a reference text. We used the SYSTRAN generic Pure Neural Server (Crego et al., 2016). We show how our scoring system (dots) compares to the official evaluation system (crosses) in Figure 2. We tend to be less generous for Exact-Match Accuracy and more optimistic for Window Overlap Accuracy (with n=3). It should be noted that our reference translation, although mostly accurate, also presents some problems. These occur mainly in the incomplete out-of-context segments related to patient symptom descriptions, many of which are also ungrammatical (ie segment 4). Table 1 recaps the scores we obtained for all the models we produced. For the models we submitted in time, as could be expected, the model trained with Common Crawl and the terminological resources (+ Term in our table) got better scores than Europarl supplemented with the terminological resources. For our in-house evaluation, we tested the translations produced by these models as well, so that we could

[4] https://commoncrawl.org/

[5] Pipelines are Hugging Face abstractions for NLP tasks that automatically select the 'correct' model architecture and all the related components (such as the tokenizer) required to make a prediction

| submitted models | BLEU (truecased) | Exact-Match Accuracy | Window Overlap Accuracy (n=2) | Window Overlap Accuracy (n=3) | 1-TERm Score | COMET |
|---|---|---|---|---|---|---|
| Common Crawl + Term | 40.02 | 0.871 | 0.296 | 0.296 | 0.507 | 0.596 |
| Europarl + Term | 34.93 | 0.795 | 0.275 | 0.267 | 0.495 | 0.296 |
| Europarl (baseline) | 33.59 | 0.640 | 0.248 | 0.241 | 0.480 | 0.212 |
| **in-house scores** | **sacreBLEU** | **Exact-Match Accuracy** | **Window Overlap Accuracy (n=2)** | **Window Overlap Accuracy (n=3)** | **1-TERm Score** | **1-TER score** |
| Hugging Face | **32.21** | 0.73 | 0.32 | **0.324** | 0.36 | 0.37 |
| mBART | 30.46 | 0.707 | 0.296 | 0.294 | 0.35 | 0.36 |
| Common Crawl + Term | 28.50 | **0.77** | 0.299 | 0.306 | 0.30 | 0.308 |
| Europarl + Term | 23.74 | 0.68 | 0.258 | 0.256 | 0.293 | 0.303 |
| Europarl (baseline) | 17.98 | 0.53 | 0.18 | 0.17 | 0.24 | 0.25 |
| Europarl (fine-tuning) | 26.19 | 0.68 | 0.279 | 0.278 | 0.278 | 0.287 |
| joeyNMT (Europarl) | 4.67 | 0.16 | 0.039 | 0.034 | 0.045 | 0.064 |

Table 1: Summary of our official and home-made scores for our models

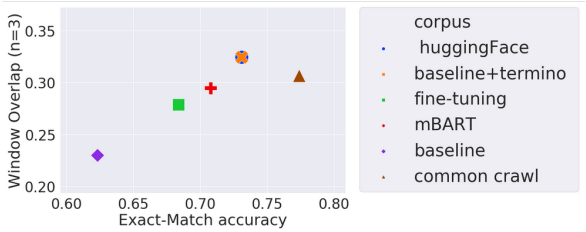


Figure 2: Comparison of the scores for the three SPECTRANS models submitted)

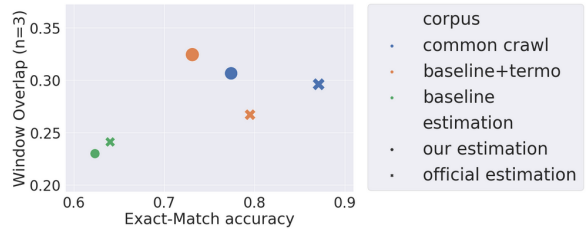


Figure 3: Comparison of the best models acccording to our scores)

compare them to the translations produced by the pre-trained models (`Hugging Face` and `mBART`. The latter did better for sacreBLEU and Window Overlap Accuracy (n=3) but probably having seen the terminological resources in the training data gave an edge for Exact-Match Accuracy to our model trained with Common Crawl and the terminological resources.

## 5 Discussion

### 5.1 Variability Across Text Genres

The benefit of our recreation of the evaluation script is that it allowed us to compute the terminology scores for 1,430 segments. We grouped the different sections of the test data according to text genres, in fashion similar to (Anastasopoulos et al., 2020). We distinguished 5 groups of texts and the variability of the BLEU scores across these text genres can be seen on Figure 4. This variability across text genres can also be seen for some other metrics, such as Window Overlap accuracy (with n=3) (see Figure 5).

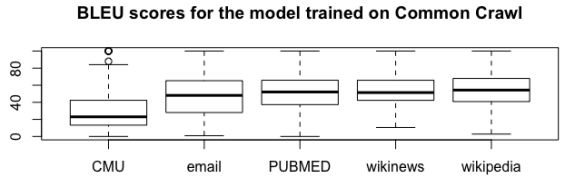


Figure 4: BLEU scores and text genres (Common Crawl training)

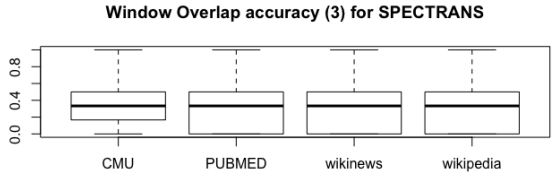


Figure 5: Variability of Window Overlap accuracy (n=3) across text genres

Overall, it is likely that our results could have been better if we had used alternative testing sets rather than using part of the reference corpora as testing sets.

### 5.2 Alternative qualitative terminological analysis

This subsection discusses the error analysis in terminology from a qualitative point of view.

For CNPs not included in the terminology dataset such as *chest pain*, the system deploys various avoidance strategies ranging from anatomic approximations (segment 20: *mal de coeur*) to omission (segment 8: *Et cette douleur est-elle bien réelle?*) to unlucky guesses (segment 2: *maux de mer*) to idiomatic expressions (segment 18: *C'est bien là que le bât blesse*). For less formal descriptions of similar symptoms where the actual term does not appear in the source text, the output

| type | segments | acronyms | terms | acronym/segment | terms/segment |
|---|---|---|---|---|---|
| CMU | 104 | 0 | 71 | 0.000 | 0.683 |
| PUBMED | 676 | 465 | 622 | 0.688 | 0.920 |
| wikinews | 67 | 11 | 25 | 0.164 | 0.373 |
| email | 98 | 7 | 7 | 0.071 | 0.071 |
| wikipedia | 1,155 | 315 | 929 | 0.273 | 0.804 |

Table 2: Distribution of acronyms in the text data

ranges from gibberish and hallucinations (segment 25: *c'est comme si la grenna est écrasée* to soaring lyricism (segment 97: *C'est la peine que j'ai sur le cœur*). When confronted with an unorganized list of terms such as the one in segment 30 (*anyone in the family have a heart problem heart disease heart attack high cholesterol high blood pressure*), most of which are not included in the dataset, the system valiantly tries to make sense of it by turning it into a complete sentence: *Quiconque au sein de la famille est confronté à un problème cardiaque, s'attaque à la pression sanguine élevée en raison de la forte pression sanguine*.

For key terminology around Covid-19, the preferred option in the output is the masculine form (*le/du/au Covid*: 127 occurrences) that is also massively present in the terminology dataset, whereas the feminine *la Covid* only appears 9 times in the output. Interestingly, in only one of these occurrences (segment 2124) does the feminine form appear within the CNP recorded in the dataset (*virus de la COVID-19*). In the other segments, it appears as a translation for the simple term COVID-19, which is in the dataset invariably associated with the masculine form when the gender is specified.

For the compound key term (56) *coronavirus disease*, different solutions appear in the output alongside the proposed translation from the dataset (*maladie du coronavirus*). One erroneous solution in our output is *maladie des coronavirus*. The plural form is problematic, as several coronaviruses exist indeed and most of them are linked with the common cold, with presents a very different picture from the illness provoked by the new coronavirus having emerged in 2019. An interesting solution appears in our output for segment 186:

[EN] The outbreak of Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2)

[FR] L'apparition de la maladie liée au coronavirus 2019 (COVID-19), causée par les syndromes respiratoires aigus sévères (SARS) Le coronavirus 2 (SARS-CoV-2)

The proposed translation, i.e. *la maladie liée au coronavirus 2019 (COVID-19)*, is actually a better choice than the one included in the dataset. The system seems to have achieved this translation by linking *disease* and *illness*, as the translation for *coronavirus disease* appears to draw from that given for *covid19 illness* in the terminology dataset. For the second CNP in this segment, *severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2)*, however, the proposed translation is less accurate, specifically in terms of syntax. This example also contains one of the few occurrences of the short form *SARS-CoV-2* in our output (16 in total, most with no article). The preferred option in our output is the translated acronym *SRAS-CoV-2*, with 153 occurrences, of which 143 also have a definite article (*le/du/au*).

### 5.3 Presence of acronyms in the terminological data

Medical terms in each segment involve two forms: acronyms and fully spelled form. The semantic fields covered by these terms include medical products ("face masks," "vaccine"), biochemical elements ("virus"), diseases ("COVID", "SARS"), as well as public health practices ("quarantine"), organizations ("WHO") and phenomena ("outbreak"). For any segment that contains at least one medical term of either form, the term count of the corresponding form is set to 1 for the segment. Counts and ratios per segment for each of the five types of documents are calculated. It can be observed from the above table that type PUBMED has the highest ratio per segment for either form of medical terms (0.688 for acronym and 0.920 for normal form), while type EMAIL has very low ratios especially for normally spelled form (0.071). In terms of medical term density, differences among these types of documents are therefore distinct. Table 2 sums up our findings in terms of the presence of acronyms

### 5.4 Terminology better at inference time?

We entered the challenge following the track for using the terminological resources at training time. We nevertheless did a background check on the possibilities of using the provided dataset at inference time. We plan to experiment the SYSTRAN Model studio functionalities to test the performance of using the terminology resource at inference time.

### 5.5 A Case for Onto-terminology?

The terminology provided for this task was unstructured, contrary to existing ontologies for medical English. Taking advantage of ontology-oriented programming in Python as implemented in Owlready (Lamy, 2017), it is tempting to consider potential implementations of onto-terminology in python-based neural translation toolkits. Biomedical ontologies have a record of established terminologies. One of the added benefits of this line of investigation is that we could not only test the gains of a structured ontology at training time but we could try to implement sanity checks at inference time to ensure the quality of the terminology by making sure the position of the terms in the output is consistent with the hierarchy in the ontology.

## 6 Conclusion

This paper presents the SPECTRANS system description for the WMT21 Terminology Shared Task. We participated in the English-to-French task, using the terminology resources at training time. Though English–French is a language-pair with many linguistic resources, we only used the data provided by the organisers. Given the novel evaluation of terminology provided for this task, we not only aimed to build a translation system for the competition, but also to provide a critical angle on the task and on its evaluation. For the MT system, we applied a variety of strategies, toolkit comparison, data augmentation and fine-tuning. Though we did not experience catastrophic forgetting, our fine-tuning did less well in the terminology metrics, probably because we were not able to update the dictionary. We obtained the best scores for the models we submitted with a model trained with Common Crawl supplemented with the terminology resource. The translations produced with pretrained models competed in terms of terminology scores, did better for sacreBLEU, especially for the translation of PUBMED, but proved less robust for the translation of the patient-doctor interactions of the CMU section of the testing data.

For the analysis of the terminology, we discussed the role of complex noun phrases and initialisms. Our contribution mostly lies in the critical analysis of the terminological input and of the evaluation script. This allowed us to raise the issue of the role of acronyms in the terminology, the importance of complex NPs (and the correlative interest of the Window Overlap Accuracy with n=3, more likely to capture complex NPs than Window Overlap Accuracy when n= 2) as well as the importance of text genres.

## Acknowledgements

The SPECTRANS project is funded under the 2020 émergence research project, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris). Lichao Zhu and Guillaume Wisniewski collaborate in this paper in the ambit of a Regional innovation programme, under the Ile-de-France DIM RFSI 2020 funding programme NeuroViz. This publication has emanated from research supported in part by a 2021 research equipment grant from the Scientific Platforms and Equipment Committee (PAPTAN project) and Masters students Internship grants to Bilal Faye and Zong-You Ke with the financial support of data intelligence institute of Paris (diip), both under ANR Grant Number ANR-18-IDEX-0001 (Financement IdEx Université de Paris).

## References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Md Mahfuz ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021. On the evaluation of machine translation for terminology consistency.

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*.

Philipp Koehn et al. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

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.

Natalie Kübler, Alexandra Mestivier, and Mojca Pecman. 2018. Teaching specialised translation through corpus linguistics: quality assessment and methodology evaluation by experimental approach. *META : Journal des traducteurs / Meta: Translators' Journal*, 63(3):806–824.

Natalie Kübler, Alexandra Mestivier, and Mojca Pecman. 2021. Using comparable corpora for translating and post-editing complex noun phrases in specialised texts: Insights from english-to-french in specialised translation. *S. Granger and M-A. Lefer (eds.), Extending the scope of corpus-based translation studies*.

Jean-Baptiste Lamy. 2017. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*, 80:11–28.

François Maniez. 2017. Evaluation des récentes avancées de la traduction automatique : le cas des adjectifs composés formés à partir d'un participe passé en anglais de spécialité. *Asp*, 72:29–48.

Alexandra Mestivier-Volanschi. 2015. Productivity and diachronic evolution of adjectival and participial compound pre-modifiers in english for specific purposes. *Fachsprache*, XXXVII(1-2):2–23.

Mojca Pecman and Natalie Kübler. 2011. Artes: an online lexical database for research and teaching in specialized translation and communication. In *Proceedings of the First International Workshop on Lexical Resources*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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

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, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.