Translating scientific abstracts in the bio-medical domain with structure-aware models

Sadaf Abdul Rauf, François Yvon

Translating Scientific Abstracts in the Bio-Medical Domain with Structure-Aware Models*

Sadaf Abdul Rauf[a,*], François Yvon[b]

[a] *Fatima Jinnah Women University, The Mall, 46000, Rawalpindi, Pakistan*
[b] *Sorbonne Université, CNRS, ISIR, 4 Place Jussieu, 75 0105, Paris, France*

Abstract

Machine Translation (MT) technologies have improved in many ways and generate usable outputs for a growing number of domains and language pairs. Yet, most sentence based MT systems struggle with contextual dependencies, processing small chunks of texts, typically sentences, in isolation from their textual context. This is likely to cause systematic errors or inconsistencies when processing long documents. While various attempts are made to handle extended contexts in translation, the relevance of these contextual cues, especially those related to the structural organization, and the extent to which they affect translation quality remains an under explored area. In this work, we explore ways to take these structural aspects into account, by integrating document structure as an extra conditioning context. Our experiments on biomedical abstracts, which are usually structured in a rigid way, suggest that this type of structural information can be useful for MT and document structure prediction. We also present in detail the impact of structural information on MT output and assess the degree to which structural information can be learned from the data.

Keywords: Neural Machine Translation, Document-level Machine Translation, Bio-medical Natural Language Processing

*This paper is published as : `https://doi.org/10.1016/j.csl.2024.101623`. Most of this work was performed while the authors were affiliated with LIMSI-CNRS (Orsay, France).

*Corresponding author.

Email addresses: `sadaf.abdulrauf@gmail.com` (Sadaf Abdul Rauf), `yvon@isir.upmc.fr` (François Yvon)

1. Introduction

Neural Machine Translation (NMT) systems have progressed remarkably well in the last decade with some claims of reaching human parity [1, 2] but much remains to be studied about the underlying learning of NMT systems. While substantial developments have been made towards integrating a document-level context [3, 4, 5, 6], the exploration of these contextual cues and the extent to which they affect translation quality remains an under explored area [7]. Translation depends upon multiple sources of dependencies between sentences in a document, that may be due to semantic, stylistic, discursive or pragmatic issues. Moreover, complex documents are also structured in parts, subparts and sections, and depending on their position within this global structure, sentences may convey varying pragmatic functions. Titles, introductory or conclusion statements, for instance, are typically thought to be playing a distinguished role and to contain the most important elements of content [8].

Sections in a document often follow distinct writing styles and lexical preferences. For instance, an analysis of 50 most frequent words from the medical abstracts used in our experiments (see Section 4.1) showed that *Titles* had the most domain specific content words; the word *"however"* was in the most frequent words for *Introduction* and *Conclusion* only; *Abstract*, *Results* and *Conclusion* had the most overlap of lexical space. Grammatically, sentences in the *Results* and *Materials and Methods* sections were majorly in past tense, whereas *Conclusion* sentences were in present continuous as well as past tense. *Objectives* are mostly written in present tense with a few sentences in past tense and so on. As sections in a document have their own grammatical and lexical blueprints, which must be reflected in their textual content, we hypothesize that informing MT systems about sentences' placement in a document may help improve the overall performance.

For a translation system, the capacity to model the context may notably improve certain translation decisions, e.g. a better or most consistent lexical choice [9], or a better translation of anaphoric pronouns [3, 5]. The research

efforts to augment NMT models with contextual or document-level information may be broadly categorized as *data-* vs *model-*centric approaches, where *data-centric* approaches focus on increasing the scope of the input to include more source and/or target texts [10, 11, 12, 13], while *model-centric* approaches aim to adjust the training architecture and objective [14, 15, 16]. Recent surveys of this subdomain of NMT are in [6, 17].

In this work, we study whether also incorporating *structural information* can help improve the MT choices and thus positively impact translation quality. Note that the structural information considered here is related to the pragmatic organization of the documents, where sectioning helps to organize the discourse into coherent subparts, each aiming to fulfill a specific goal. This is in contrast with other attempts to translate structured documents [18, 19] which focus on integrating syntactic markers of the structure or [20], where structural organization identifies thematic changes. In our *data-centric* approach, contextual and structural information are injected for each sentence based on a hierarchy of special tokens and does not require any change in the model. While simple, this technique allows us to study whether learning coherent, section-specific, translation styles can contribute to improving the overall performance.

As tags can be used in multiple ways, we further explore the differences in tagging the source or the target side, which also allows us to evaluate how reliably the structure information can be learned automatically from the data. We systematically study the impact of each tag on metric score and the translation output in conjunction with human analysis of section wise lexical preferences versus the MT models section prediction capabilities.

Our experiments focus on abstracts of scientific texts in the bio-medical domain [21, 22], where document structure information is often available. For this target genre, texts are expected to follow a standardized structure comprising typical subsections of one to five lines, known as the IMRaD structure (for **I**ntroduction, **M**ethod, **R**esults, and, **D**iscussion) [23]. As we discuss in Section 3.1, preparing the data for these experiments has been challenging and a significant outcome of our work will be a new dataset annotated with document

structure, that we will release to the community [1].

In summary, the main research questions that we address in this work are:

1. **[RQ1]** Does incorporating document structure information in text help improve machine translation performance?
2. **[RQ2]** What is the most effective way to incorporate this auxiliary information?
3. **[RQ3]** When the structure is not observed, how well can it be automatically predicted? What will be the effect of prediction errors?

This paper is organized as follows: we start with a brief overview of Neural Machine Translation models (section 2), followed by an elaboration of our approach and architecture in sections 3 and 4. Results based on automatic metric scores are detailed in sections 5 and 6, whereas the analysis of a human evaluation is presented in 6.3. The paper concludes with an overview of related works (Section 7) and a general discussion in Section 8.

2. Neural Machine Translation

2.1. Probabilistic encoder-decoder models

In this section, we briefly introduce the main concepts of neural machine translation (NMT), with an emphasis on concepts whose understanding is important for accurately describing methods that handle context. A much more complete presentation of NMT is in [24, 25].

The general principle of all MT architectures is to generate the best possible target translation $\mathbf{e}= e_1 \dots e_T$ of the input source sentence $\mathbf{f}=f_1 \dots f_S$, according to a probabilistic decision rule:

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(\mathbf{e}|\mathbf{f}), \tag{1}$$

where the model parameters ($\boldsymbol{\theta}$) are estimated from sentence-aligned parallel corpora. Learning such a distribution is unrealistic if one considers complete

[1] `https://github.com/fyvo/WMT-Biomed-Test`

sentences as the random variable in the model. However, neural implementations of this probabilistic model enable to directly factorize this distribution as:

$$p_{\boldsymbol{\theta}}(\mathbf{e}|\mathbf{f}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f}), \tag{2}$$

without having to resort to conditional independence hypotheses (eg. Markovian dependencies).

Equation (2) simply states that the probability of each target word is generated conditioned on the current prefix of the target sentence ($\mathbf{e}_{<t} = e_1 \dots e_{t-1}$) and on the entire source sentence ($\mathbf{f}$). The manipulation of such distributions is made possible by transforming discrete contexts ($\mathbf{e}_{<t}, \mathbf{f}$) and words $\mathbf{e}_t$ in the vocabulary into continuous representation spaces. This means that each word $\mathbf{f}_t$ (and accordingly each context or each substring) is associated with a large numerical vector representing the useful information about that word or context.

Factorization (2) suggests that the program (1) can be solved approximately by generating words from left to right according to the following greedy procedure:

$$p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{f}) = \operatorname{argmax} \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f}). \tag{3}$$

More effective search procedures can also be entertained (see [26] for a review).

The estimation of $\boldsymbol{\theta}$ is performed by maximizing the conditional log-likelihood (or cross-entropy) $\sum_{t=1}^{T} \log p_\theta(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f})$ accumulated over a large set of sentences, resulting in a complex optimization program that is typically solved in an approximate manner by generic numerical optimization algorithms.

The main neural architectures are mainly distinguished according to the way the conditioning context is encoded. The Transformer architecture [27] replaces the recurrent component (in the source and target) introduced in [28] by *generalized attentional modules*, while retaining the source-target cross-attention component of early RNN-based architectures. This change makes all previous positions $t-1, t-2 \dots 1$ equally important in the selection of the current word (and likewise for source representations). It also enables to process all the target tokens in parallel during training, causing vast decrease in training time.

Decoding must still be carried out from left to right, since each word already generated conditions the generation of future words.

2.2. Monolingual and Bilingual Attention Modules

In a nutshell, the Transformer architecture transforms a *structured context* (a sequence of tokens, but this can also be a tree or a graph) into a single numerical vector. The core operation in this transformation is the iterative computation of each individual token's representation based on their similarity to other tokens in the context. Denoting $H^{l-1} = [h_1^l, \ldots, h_T^l]$ the $(T \times d)$ matrix representing a context of length T at the input of layer l, the representation $\tilde{h}_i^{kl}$ for token i is computed by *attention head kl* as:

$$\tilde{h}_i^{kl} = \text{softmax}(\frac{h_i^{l-1} Q^{kl} [H^{l-1} K^{kl}]^T}{\sqrt{d_k}}) H^{l-1} V^{kl}, \tag{4}$$

,

with Q^{kl}, K^{kl}, V^{kl} the parameter matrices associated to this head and layer, d the model dimension, and d_k the size of each of the K heads ($d_k = d/K$). In this model, H_0 contains the lexical and positional embeddings. The output of these k computations are then concatenated and passed through a feed-forward (FFN) layer with ReLU activation; each of these steps also includes a summation with H^{l-1} and a layer normalization. In equations:

$$\begin{aligned}
\tilde{H}^{kl} &= \text{Attn}(H^{l-1}Q, H^{l-1}K^{kl}, H^{l-1}V^{kl}) \\
H^l &= \text{LayerNorm}(H^{l-1} + [\tilde{H}^{1l}, \ldots, \tilde{H}^{Kl}]) \\
H^l &= \text{LayerNorm}(H^l + \text{FFN}(H^l)),
\end{aligned}$$

where Attn denotes the computation expressed by equation (4).

When used for MT, the self-attention mechanism is computed on the source side *on the full observed sentence*. On the target side, the self-attention is restricted to the previous words; an additional *cross-attention* module combines at each layer the decoder-side representations with the encoder representations output by the top layer H^K: this way, information may flow from source to target, but not in the reverse direction.

3. Methods: Translating with Document Structure

We propose to introduce document structure information at the level of each sentence through a system of hierarchical tags. Tags have been widely used in NMT to incorporate additional discrete conditioning factors in (1), often for issues quite distinct from those caused by long-range dependency, e.g. to control the output language, the domain, or the level of formality and politeness [29, 30, 31]. Tags also provide us with a simple and effective way to take structure into account.

3.1. Tagging sections within documents

3.1.1. The IMRaD scheme

Our starting point is that scientific documents in our target domain often follow a rigid structure, known as the IMRaD structure [23]. This means that most abstracts will comprise one or several lines for the following sections: ***I****ntroduction*, ***M****ethod*, ***R****esults*, and, ***D****iscussion* (see example in Figure 1), where we also consider the title line to be part of the structure.

Note that this structure can be overt and introduced by appropriate headings, or remain covert. In the latter case, the document structure cannot be immediately recovered from the text. Also note that variations in the organization and in the heading labels can be observed depending on the subdomains and publication venues.

Our main hypothesis is that each of these sections can be characterized with specific terms and phraseological patterns: in our experiments, we thus explore how to use this information in NMT and measure the correlated impact. We notably expect that by informing the system with sub-document information, it will better translate the typical style and phraseology of sentences occurring in each part.

For this purpose, we identified in our data all the abstracts that were conforming to this basic structure and worked to make this structure as explicit and standardized as possible. This notably implied to normalize the main headings,

Title:	Comparison of two azithromycin distribution strategies for controlling trachoma in Nepal.
OBJECTIVE:	The study compares the effectiveness of two strategies for distributing azithromycin in an area with mild-to-moderate active trachoma in Nepal.
METHODS:	The two strategies investigated were the use of azithromycin for 1) mass treatment of all children, or 2) targeted treatment of only those children who were found to be clinically active, as well as all members of their household.
FINDINGS:	Mass treatment of children was slightly more effective in terms of decreasing the prevalence of clinically active trachoma (estimated by clinical examination) and of chlamydial infection (estimated by DNA amplification tests), although neither result was statistically significant.
CONCLUSION:	Both strategies appeared to be effective in reducing the prevalence of clinically active trachoma and infection six months after the treatment. Antibiotic treatment reduced the prevalence of chlamydial infection more than it did the level of clinically active trachoma.

Figure 1: A structured abstract with an overt organization, extracted from Scielo (fr-en) test set, a bio-medical corpus. In this example, the *Introduction* is missing, and an*Objective* part is present; the *Result* part is labeled 'FINDINGS'.

as there was a large degree of variance in subheadings and structures across corpora (see § 4.1.3). To incorporate the standard IMRaD format, we thus mapped each subheading to the corresponding IMRaD label using a system of tags displayed in Table 1. For instance, *Subjects and Methods* was mapped onto *Material and Methods* (<MaM>), *Pedagogical objectives* may be replaced with *Objectives* (<OBJ>), *Novel finding* with *Results* (<RES>), and so on. We also found it appropriate to merge some sections, such as *Discussion*, which was merged with *Conclusion*.

Document level corpora (see § 4.1.3) were first retrieved in XML format and structured into sub-sections based on subheading information. Each section was split[2] into sentences and sentence aligned using Microsoft bilingual aligner [32].

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

The identification and standardization of the subheading information was a tedious process, involving a lot of rule-based processing and human intervention to take the variability of sub-headings into account. In order to reconstruct the fully parallel versions with subheadings, we also had to reinsert explicit headings in the source or target files when they were missing. Also, note that this information was not available for all biomedical abstracts. In this case, we tagged every line but the title with a generic tag (<ABS>).

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

Table 1: Standardized section heading and the corresponding tags

3.2. A hierarchy of tags

Building state-of-the-art NMT systems for abstracts in the biomedical domain requires to consider multiple sources of parallel data, opportunistically collecting texts from a variety of genres and domains [33]. Our own training data, presented in Section 4.1, is accordingly very diverse, comprising in-domain parallel and out-of-domain parallel corpora, texts automatically retrieved parallel collections, as well in-domain monolingual data that is automatically back-translated. Some are made of lists of isolated sentences, while others retain the document information. Even within the in-domain data, some texts precisely match the genre of the test set (scientific abstracts), whereas others can be quite remote (eg. regulatory documents from the EMA[3]). This means that for most of our training data, the document structure will often be either unknown, or will not correspond to the scientific genre and will not obey the IMRaD pattern.

In order to reflect this diversity and provide tags for all parallel sentences in our corpus, we extended our tagging scheme with two additional levels of tagging. We then use hierarchical tags of the form (<1> <2> <3>), to incorporate a

[3]European Medicines Agency, formerly known as EMEA.

<M> <SCI> <H1>	Comparison of two azithromycin distribution strategies for controlling trachoma in Nepal.
<M> <SCI> <OBJ>	OBJECTIVE
<M> <SCI> <OBJ>	The study compares the effectiveness of two strategies for distributing azithromycin in an area with (...)
<M> <SCI> <MaM>	The two strategies investigated were the use of azithromycin for 1) mass treatment of all children, or 2) (...)
<M> <SCI> <RES>	FINDINGS
<M> <SCI> <RES>	Mass treatment of children was slightly more effective in terms of decreasing the prevalence of clinically (...)
<M> <SCI> <CON>	CONCLUSION
<M> <SCI> <CON>	Both strategies appeared to be effective in reducing the prevalence of clinically active trachoma and (...)
<M> <SCI> <CON>	Antibiotic treatment reduced the prevalence of chlamydial infection more than it did the level of clinically (...)

Figure 2: Examples of fully tagged sentences from Scielo, a bio-medical corpus. Tags appear as prefixes in the source sentence.

multi-level information about the domain, corpus and section for each sentence. The first level of tags distinguishes between out-of-domain data (<G>), and in-domain data (<M>). The second level of tag aims to distinguish between data sources, hence the use of one dedicated tag for each corpus, except for data collected with Information Retrieval (IR) techniques and monolingual data, which are respectively tagged with <IR> and <BT> [34, 35]. Finally, the third tag either corresponds to the structural information as described above, when it is available; otherwise all remaining sentences from other corpora were simply tagged with an "unspecified subheading" label (<US>).

Fully tagged sentences sampled from the Scielo test document displayed on Figure 1 are reproduced on Figure 2. In this example, tags are prefixed to the source sentence: the first tag (<M>) specifies the bio-medical domain, the second tag identifies the corpus (<SCI>), and the third tag corresponds to the subsection of each sentence.

3.3. Using tags in MT

Having defined a tagging scheme, the next question concerns its use in machine translation. A first alternative is to use tags as words and insert them as prefix on the source side. This assumes that tags are always known prior to translation, and will have the effect to differentiate the contextual representations of source words, enabling the system to disambiguate the translation of polysemous words, whenever this is beneficial for the training loss. The same effect will be observed on the target side, thanks to the cross-lingual attention mechanism alluded to above. It is further possible to explicitly amplify this effect by injecting the tag information into the input representation of every token (eg. [30, 36] for domain information, or [37] for language information).

The alternative is to inject the tags as a prefix of the *target* sentence. If we assume that tags are always observed, the expected effect is roughly similar to having tags on the source side. The main difference is that source token representations no longer depend on the tag value. We have not studied this approach, but have instead resorted to another way to use target side tags: to process them as if they were regular target words. The system will then be trained to predict their correct values using only the source side information (and the previous tag(s)). In this setting, source side representations will be adjusted to correctly generate tags and to extract features that are useful for this prediction [38]. With forced decoding [39], this approach can still be used when tags are observed. In addition, it can also be used to generate tag-dependent translations, for unstructured abstracts, with unobserved tags. These two schemes are considered in our experiments and help to better analyse the usefulness of the structural tags. Similar analyses, for a variety of tags, are in [40, 41, 42]. They suggest that when tags are observed, having them on the source side yields better results. Note that these studies only consider one level of tagging, where we consider three.

To sum up, three scenarios will be contrasted below: (i) observed source side tags, (ii) observed target side tags (forced decoding), (iii) predicted target side tags: (i) and (ii)-(iii) differ in the training scheme, while (ii) and (iii) only differ

Corpus	Wrds (M)			Corpus	Wrds (M)		
	English	French	Sents.		English	French	Sent.
In Domain				Monolingual			
Edp	0.04	0.04	3.3 K	Med_Fr	1.34	1.63	0.06 M
Medline titles	5.97	6.43	0.62 M	IsTex_Fr	6.92	7.84	0.42M
Medline abstracts	1.23	1.44	57.6 K	Lissa_Fr	8.79	7.70	0.33 M
Scielo	0.17	0.21	12 K	Med_En	3.40	4.02	0.22M
Cochrane-Reference	2.23	2.74	0.12 M	Development			
Cochrane-PE	0.43	0.53	25.6 K	Medline 18	5.7K	6.9K	265
Cochrane-GooglePE	0.63	0.77	37.8 K	Medline 19	9.8K	12.4K	537
Total <Section>			0.87 M	Test			
Ufal	89.5	100.3	2.62 M	Scielo (wmt16)	0.14	0.17	6475
Taus	20.1	23.2	0.88 M	Medline 20	25K	28K	997
Mlia	19.0	23.0	1.0M	Newstest	93K	0.1	3003
IR Retrieved	13.2	14.7	3.6M	**Out Domain**			
Total-in-dom<US>			8.2 M	Out-domain<US>	1139	1292	35M

Table 2: Data sources used in our study. Test sets of both directions with only document information were combined.

in the way inference is performed.

4. Experimental Setup

4.1. Corpora and Pre-processing

We train our baseline systems on a collection of in-domain biomedical texts as well as out-of-domain parallel corpus, and use standard benchmarks from past studies for development and tests. Details are in Table 2.

4.1.1. Parallel corpora

We gathered parallel and monolingual corpora available for English-French in the biomedical domain. The former included the biomedical texts provided by the WMT'20 organizers: Edp, Medline abstracts and titles [43], Scielo [44] and the Ufal Medical corpus[4] consisting of Cesta, Ecdc, Emea (OpenSubtitles), PatTR Medical and Subtitles. Other sources of in-domain data are the

[4] `https://ufal.mff.cuni.cz/ufal_medical_corpus`

Cochrane bilingual parallel corpus [45][5], the Taus Corona Crisis corpus[6] and the Mlia Covid corpus[7]. We also included additional in-domain data selected using Information Retrieval (IR) techniques from general domain corpora including News-Commentary, Books and Wikipedia corpus obtained from the Open Parallel Corpus (OPUS) [46]. These were selected using the data selection scheme described in [47]. Medline titles were used as queries to find relevant sentences. We included the `2-best` sentences returned from the IR pipeline into our additional corpus [48].

Our out-of-domain corpora include the parallel data provided by the WMT'14 campaign for French-English: Gigafr-en, Common Crawl, Europarl, News Commentary and the UN corpora.

For development purposes, we used Medline testsets for WMT'18 [49] and 19 [50] bio-medical MT task, while Medline 20, Scielo (WMT'16) and Edp (WMT'17) were used as internal test data.[8] Test sets for both directions were combined to get larger test sets and are used to evaluate systems in both directions. We only used the parts with document structure, i.e. sentences with <ABS> were not included. The combined test set for Scielo and Medline had $8,805$ and 1315 sentences each, after removing the abstract sentences we get the test sets with 6475 and 997 sentences for Scielo and Medline20 respectively.

4.1.2. Monolingual sources

The backtranslation of monolingual sources has often been effectively used to cater for parallel corpus shortage in the bio-medical domain [51, 33]. We also augment our training data with backtranslations using both French and English in-domain monolingual corpora.

Supplementary French data from three monolingual sources were collected

[5] `https://github.com/fyvo/CochraneTranslations/`

[6] `https://md.taus.net/corona`

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

[8] These test sets were sentence-aligned with in-house tools and can be retrieved from `https://github.com/fyvo/WMT-Biomed-Test`.

from public archives: 1) abstracts of medical papers published by Elsevier from the Lissa portal;[9] 2) abstracts from the national ISTEX archive;[10] 3) a collection of research articles collected from various sources[11] [52] henceforth referred to as Med_Fr. These documents were automatically translated into English with an NMT system trained only on bio-medical corpora, with a reported BLEU score of 33.6 on Medline20 testset.

The English side of the Medline German and Spanish corpora distributed for WMT/Biomed 2021[12] is used as supplementary English data for backtranslation. Duplicate documents were removed based on the document id information. For these abstracts, the internal structure of documents is often available and has been tagged as for the parallel data. These texts were then split into sentences[13] and translated into French using an NMT system trained on all bio-medical corpora with a BLEU score of 36.4 on Medline20 testset.

4.1.3. Structured sources

The document structure is only available for a subset of the training data corresponding to the following sources:

1. *Medline and Scielo:* These included both abstracts with and without subheadings, sometimes also missing a title line. There were a total of 189 subheadings in these corpora, including variations in case, word order and morphology. Examples subheadings include: *Presenting Concerns of the Patient*, *Sources of Information*, *Novel finding*, *Study Selection*, *PURPOSE OF REVIEW*, *Purpose of Review*, *Purpose of review* etc.
2. *Edp:* These documents included abstracts with and without subheadings, with a majority having titles. Total 45 subheadings with variations in case and word order were identified such as: *Case report*, *Observation*, *Subjects and Methods*, *Commentary*, *Pedagogical objectives*, etc.

[9] https://www.lissa.fr/dc/#env=lissa
[10] https://www.istex.fr/
[11] https://crtt.univ-lyon2.fr/les-corpus-medicaux-du-crtt-613310.kjsp
[12] http://statmt.org/2021
[13] https://pypi.org/project/sentence-splitter/

3. *Cochrane:* This corpus contains very structured summaries with 10 distinct subheadings according to the Cochrane nomenclature. These included: *abs selection criteria*, *abs search strategy*, *abs data collection*, *summary title* etc. [14]

For the test and development corpora, when the structure was available, a manual alignment was performed based on the sentence alignment file provided by WMT organizers.[15] Alignment operations involved inserting an empty line after each subsection, deleting unaligned sentences and merging sentences as per alignment information. Statistics regarding the numbers of structured documents in the train, test and development corpora are in Table 2.

4.2. Translation systems

Our translation systems mostly use the basic Transformer architecture [27]. They all rely on Meta's seq-2-seq library (fairseq) [53] with parameter settings borrowed from the `transformer_wmt_de_en` recipe.[16] The ReLU activation function was used in all encoder and decoder layers. We optimize with Adam [54], set up with a maximum learning rate of 0.0005 and an inverse square root decay schedule, as well as 4000 warmup updates. All corpora are segmented into subword units using Sentence Piece [55] with a vocabulary of 32K units. These units were learned on the union of all in-domain corpora. We share the decoder input and output embedding matrices.

Models are trained with mixed precision and a batch size of 4096 tokens on 4 V100 GPUs for 300k updates until convergence, where convergence is measured on the development set BLEU score. Final parameters are chosen based on the best score on the development set (Medline18, 19) and the corresponding scores for that checkpoint are reported on Medline20 test set. Translations were generated using a beam size of 4. For TRGTAG systems tags were removed before

[14] Recommandations for writing Cochrane's systematic reviews are in `https://training.cochrane.org/handbook/`.

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

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

automatic evaluation. Evaluation was performed using SacreBleu [56] using BLEU [57], TER [58] and chrF [59] metrics.

For fine-tuned systems, the process starts with models converged systems. Training then resumes using a selected portion of the train data, with the same meta parameters and objective as for the base systems. In our results, names of fine-tuned systems are post-fixed with `*-ft`.

5. Document Structure Information and MT performance

To evaluate the impact of document structure information on NMT, we built systems with tags either on the source or target sides and compared them with baseline NMT systems. Scores show that incorporating tags substantially improves machine translation performance; this improvement is more prominent when tags are used on the source side to initialize the encoder input stream. These results are tabulated in Table 3 on two bio-medical test sets, `Medline20` and `Scielo` for standard MT systems (*All*), source-tagged systems (SRCTAG) and destination-tagged systems (TRGTAG). Baseline systems use all general domain and biomedical corpora totalling 41.7M sentences for EN-FR and 40.9M sentences for FR-EN. Each of these systems is then further fine-tuned on bio-medical corpora (5.7M sentences for EN-FR, and 4.9M sentences for FR-EN).

We use three MT metrics to evaluate the systems in this section. These include BLEU which uses n-gram precision and Brevity penalty, TER which gives scores based on edit operations and chrF which uses F-score statistic for character n-gram matches that a hypothesis requires to match the reference translation. Since TER is an error rate, a decrease in score signifies improvement, while for BLEU and chrF increase denotes improvement. We see that on average the three metrics agree with each other in the general trends exhibited by the systems, but BLEU is more prominent, thus we will present further analysis based on BLEU scores.

A general trend is that systems with extra document information perform substantially better than vanilla NMT systems. We find that it seems better to

	Medline20						Scielo					
	EN-FR			FR-EN			EN-FR			FR-EN		
	BLEU	TER	chrF	BLEU	TER	chrF	BLEU	TER	chrF	BLEU	TER	chrF
All	39.3	50.2	65.5	40.5	49.0	67.0	36.4	52.3	63.4	36.3	53.2	65.6
+ft	40.3	48.9	66.3	43.1	46.5	68.7	36.0	52.0	63.3	36.5	52.6	66.0
SRCTAC	40.2	49.0	66.3	40.6	48.7	67.1	36.9	52.0	63.8	37.0	51.7	66.0
+ft	**41.5**	**48.4**	**66.7**	**44.0**	**45.3**	**69.0**	36.8	51.8	63.8	**38.6**	**50.1**	**66.7**
TRCTAC												
Reference	40.1	49.9	65.8	40.1	49.9	67.1	***37.0***	51.9	***63.9***	36.2	53.5	65.6
+ft	*41.0*	*48.7*	*66.5*	*43.2*	*46.3*	*68.8*	36.8	***51.6***	***63.9***	*37.7*	*51.4*	*66.3*
Predicted	39.6	50.2	65.6	40.0	50.1	67.0	36.1	52.3	63.3	35.1	54.8	65.3
+ft	40.3	49.2	66.3	42.6	46.9	68.6	35.6	52.3	63.2	35.9	53.5	65.8

Table 3: BLEU, TER and chrF scores computed on Medline20 and Scielo test sets with and without tags, where we contrast tags in source (SRCTAG), reference target tags (TRGTAG with forced decoding) and predicted target tags (TRGTAG). `+ft` identifies the corresponding fine-tuned systems. Note that for TER, a decrease in score signifies improvement, while for BLEU and chrF increase denotes improvement. The best result in each column is in bold and best score for predicted vs. reference tags are italicized .

introduce the tags on the source than on the target side, similar to [40, 41, 42]. We see an average improvement of 0.92 and 0.46 BLEU points on the source and target side tags respectively. Further, in TRGTAG setting, using reference tags is consistently better than predicting them, with more than 1 BLEU point difference between these settings for Scielo. These observations carry over for fine-tuned systems. We also see that the effect of fine-tuning is generally positive, with the exception for Scielo systems with TRGTAG for EN-FR.

5.1. Ablation Experiment

An ablation experiment was conducted in the SRCTAG setting for French-English direction to isolate the impact of document structure information. For this, we trained two models with one tag each: the first model with only domain tags, the second with only document structure (section) tags. This is in contrast to our experiments with the hierarchical tagging scheme where 3 dimensions of information are used to condition the translation output. The rationale behind this ablation experiment is to isolate the effect of other dimensions.

The section and domain tagged systems had a BLEU score of 36.7 and

1 - src	Compte tenu du manque de renseignements concernant ce modèle modifié, nous avons procédé à une étude pour en déterminer **la survie et le rendement à court et à moyen terme.**
<INT>	Given the lack of information about this modified model, we conducted a study to determine **short- and medium-term survival and performance.**
<M>	Given the lack of information on this modified model, we conducted a study to determine **its survival and performance in the short and medium term.**
2 - src	S'il existe de nombreuses approches **chiropratiques**, deux types de **chiropracteurs** peuvent-être identifiés; ceux s'intéressant aux troubles musculo-squelettiques et **ceux souhaitant prendre en charge aussi des troubles non musculo-squelettiques.**
<INT>	While there are many **chiropractor** approaches, two types of **chiropractors** may be identified; those interested in musculoskeletal disorders and those **wishing to also manage non-mechanic disorders.**
<M>	If there are many **cheropractical** approaches, two types of **cheropractors** may be identified; those interested in musculoskeletal disorders and those **interested in the management of non-mechanic disorders as well.**
3 - src	**Il est préoccupant de constater que** les étudiants qui adhèrent au modèle de la **subluxation** soient prêts à **intégrer** ces opinions dans leurs futures **prises en charge**; souhaitant proposer des ajustements **chiropratiques** aux patients asymptomatiques.
<CON>	**Concerningly,** students who adhere to the **subluxe** model are prepared to **integrate** these views into their future **care**; wishing to propose **chiropractical** adjustments to asymptomatic patients.
<M>	**It is worrying to note that** students who adhere to the **subluxation** model are prepared to **incorporate** these views into their future **management**; wishing to propose **cheropractical** adjustments to asymptomatic patients.
4 - src	Les symptômes courants sont des maux de tête, des troubles de la vue, des acouphènes pulsatiles et un oedème papillaire.
<INT>	Common symptoms are headaches, **visual** disorders, **perfumant psychosis**, and **hematomary** edema.
<M>	Common symptoms are headache, **vision** disorders, **puffering drowsiness**, and **facial** edema.

Table 4: Ablation experiment: Output from model trained with only document structure tags (a) and only domain tags (b).

36.2 respectively (at 360M updates). Section tagging gave an improvement of 0.5 BLEU points. We observe the impact on output in Table 4 where sentences 1-2 display a common pattern: the difference in grammatical preference between models trained with only section information vs only domain information. Sentences 3-4 highlight variance in lexical choice for the two models. The difference in average sentence length of the output of the two systems was negligible (0.26 words). From these results, we conclude that compounded effect of hierarchical tagging is more pronounced as compared to using only one level of tagging.

6. Joint Prediction and Translation

In this section, we evaluate TRGTAG systems in detail. Recall that for each sentence the model has to successively predict three tags, corresponding respectively to the domain, corpus and the section, before generating the output translation. We thus systematically evaluate and contrast (i) the accuracy of the prediction of each tag, and (ii) the impact of having a correct vs. incorrect prediction on the BLEU score. To do so, we use the systems trained with target side tags as prefix of the reference translation. This setting allows us to study how well tags can be automatically predicted from the source sentence and the resulting impact on the final translation.

For this, we force decode each test sentence four times with four prefixes of increasing size: `pre0` is the empty prefix condition, where all tags are automatically generated, while `pre3` initializes the decoder with the three reference tags; `pre1` and `pre2` denote the two intermediary settings, respectively with one and two correct tags. The former two conditions respectively correspond to decoding with predicted or reference tags in Table 3, the latter two are novel.

6.1. Using reference vs. predicted tags

In our results, we bin the sentences based on tag predictions. Bin 000 contains sentences for which that all three tags are wrongly predicted; similarly

All			EN-FR: BLEU Scores (Percentage of sentences)					
Prefix	Testsets (# Sent.)	Global Scores	000	001	100	101	110	111
Pre0		36.8	31.0 (1.5%)	-	36.7 (66.5%)	38.1 (21%)	29.2 (0.3%)	37.6 (11%)
Pre1	scielo	36.8	-	-	36.6 (68%)	37.9 (21%)	29.2 (0.3%)	37.6 (11%)
Pre2	(6475)	**37.5**	-	-	-	-	35.8 (39%)	39.6 (61%)
Pre3		**37.5**	-	-	-	-		37.5 (100%)
Pre0		41.0	44.7 (1.1%)	-	39.9 (59.5%)	44.5 (15.2%)	39.8 (7.7%)	47.2 (16.5%)
Pre1	medline20	41.0	-	-	39.8 (60%)	44.4 (15.5%)	39.8 (7.7%)	48.7 (17%)
Pre2	(997)	**41.4**	-	-	-	-	38.2 (55%)	46.0 (45%)
Pre3		40.1	-	-	-	-	-	40.1 (100%)
+fine tuning on Biomed								
Prefix	Testsets	Scores	000	001	100	101	110	111
Pre0		36.5	-	-	35.7 (49%)	38.5 (48%)	29.0 (0.5%)	39.4 (2.4%)
Pre1	scielo	36.5	-	-	35.7 (49%)	38.5 (48%)	29.0 (0.5%)	39.4 (2.4%)
Pre2	(6475)	**37.7**	-	-	-	-	35.3 (24.5%)	39.1 (75.5%)
Pre3		37.6	-	-	-	-	-	37.6 (100%)
Pre0		42.0	-	-	40.2 (45%)	45.8 (15%)	39.8 (10%)	46.2 (30%)
Pre1	medline20	42.0	-	-	40.2 (45%)	45.8 (15%)	39.8 (10%)	46.2 (30%)
Pre2	(997)	**42.4**	-	-	-	-	39.6 (49%)	45.6 (51%)
Pre3		41.9	-	-	-	-	-	41.9 (100%)
All			FR-EN:BLEU Scores/Percentage of sentences					
Prefix	Testsets	Scores	000	001	100	101	110	111
Pre0		35.5	45.8 (0.2%)	-	35.6 (89.8%)	34.5 (10%)	-	-
Pre1	scielo	35.5	-	-	35.6 (89.8%)	34.5 (10%)	-	-
Pre2	(6475)	**36.8**	-	-	-	-	33.2(39%)	39.5(61%)
Pre3		**36.8**	-	-	-	-		36.8 (100%)
Pre0		40.5	46.8 (0.4%)	-	40.4 (80%)	55.1 (2.6%)	31.5 (13%)	44.2 (4%)
Pre1	medline20	40.5	-	-	40.5(80%)	55.1 (2.6%)	31.5 (13%)	44.2 (4%)
Pre2	(997)	40.7	-	-	-	-	36.6 (60%)	47.0 (40%)
Pre3		**40.7**	-	-	-	-	-	40.7 (100%)
+fine tuning on Biomed								
Prefix	Testsets	Scores	000	001	100	101	110	111
Pre0		36.6	-	-	37.0 (88.6%)	32.5 (10%)	36.4 (0.3%)	38.0 (1.5%)
Pre1	scielo	36.6	-	-	37.0 (88.6%)	32.5 (10%)	36.4 (0.3%)	38.0 (1.5%)
Pre2	(6475)	38.7	-	-	-	-	34.9 (23%)	40.6 (77%)
Pre3		**38.8**	-	-	-	-		38.8(100%)
Pre0		43.6	-	-	42.6 (71.5%)	59.1 (4%)	35.4 (15%)	48.6 (9.5%)
Pre1	medline20	43.6	-	-	42.6 (71.5%)	59.1 (4%)	35.4 (15%)	48.6 (9.5%)
Pre2	(997)	**44.3**	-	-	-	-	39.1 (46%)	49.9 (54%)
Pre3		**44.3**	-	-	-	-	-	44.3 (100%)

Table 5: Evaluating the difference in translation quality of sentences with correctly predicted sections vs. wrongly predicted sections (Domain, Corpus, Section). 000 indicates that the three tags are wrongly predicted, 111 that they are all correct.

bin 111 contains sentences with three correct predictions, 101 those with a correct prediction for the first and third tags and so on. To measure the translation quality, BLEU scores are computed for each subset of sentences, excluding as before tags in the score computation. Condition `pre0` is when we predict all three tags, while `pre3`, `pre2` and `pre1` involve some reference tag(s). Table 5 reports the BLEU scores and percentage of sentences in each bin.

We can make the following observations that apply for the two language directions. First, it turns out that predicting the first tag (domain) is easy which also explains the small gap in BLEU scores between pre0 and pre1. Having the second tag (corpus) right is the most challenging part, with an error rate higher than 75% for the non fine-tuned systems. This is because some of our in domain corpora are very close (eg. Cochrane and Medline). It remains difficult even after fine-tuning, with an error rate way above 60%. The third tag (section headings), is much easier — assuming the other tags are correct, we achieve accuracy higher than 50% (without fine-tuning) and higher than 60% with fine-tuning. When we get the corpus tag wrong, we are likely to fail for section prediction (see 100 vs. 101), i.e. if the system errs on corpus it is likely to err also on the section tag. The main reason for this is data imbalance with just 8.7% in-domain sentences having document structure information (see Table 2).

BLEU-wise, we see overall a general improvement when we move from predicted tags (`pre0`) to correct tags (`pre3`); when the corpus is known, section prediction is relatively easy explaining why `pre2` and `pre3` are always very close, and in some cases it even seems slightly better to let the system use predicted sections headings than use the correct ones.

Structure vs No structure tags. As `pre2` and `pre3` scores are always very close, to better evaluate the impact of using reference section tags, we design another contrast and force decode all the test-sets with a 'generic' section tag instead of the actual IMRaD code. Results are in Table 6 and again contrast the effect of using source vs. target tags.

Having structure tags in source slightly improves BLEU for the fine-tuned systems on Medline (+0.6 and +0.7 respectively for EN-FR and FR-EN). All other comparisons show little, insgnificant variations. The same small variations are observed for TRGTAG systems, where we even observe that the generic tag yields a tiny improvement over the other condition. These small differences may be due to the training data distribution, as 75% and 48% sentences of Medline and Scielo respectively had <ABS> as the section tag. Overall, these results do hint at a slight improvement in BLEU when the more specific section is known, but given the small difference and the data imbalance, we abstain to make any conclusive statement.

Test set	EN-FR		FR-EN	
SRCTAG	Spec.	Gen.	Spec.	Gen.
Scielo	**37.6**	37.5	**37.6**	37.5
+ft	37.6	37.6	39.7	**39.8**
Medline20	**41.8**	41.4	41.6	41.6
+ft	**42.6**	42.0	**45.4**	44.7
TRGTAG	Spec.	Gen.	Spec.	Gen.
Scielo	37.5	37.5	**36.8**	36.7
+ft	**37.7**	37.6	**38.8**	38.6
Medline20	41.0	**41.4**	40.7	**40.8**
+ft	41.9	**42.3**	**44.3**	44.0

Table 6: Using specific or generic section tags

6.2. Predicting Subsection Tags

For this analysis, we report results on the Scielo testset, which among all the biomedical testsets, has the largest number of documents annotated with structure information (cf. Section 4.1.3). Confusion matrices for EN-FR are shown in Figure 3, and the corresponding metric scores and FR-EN confusion matrices are given in Appendix A.

Section prediction results as shown in Table 7 resonate with our human analysis of lexical and grammar usage in different sections shown in Table 8. We analysed 50 most frequent words per section along with their POS tags (using nltk). Some words are shown in Table 8 to demonstrate that the word itself is an indicator of its' most probable section and their usage is enriched in particular sections, e.g. the words like *however* and *may* exhibit higher probability to be

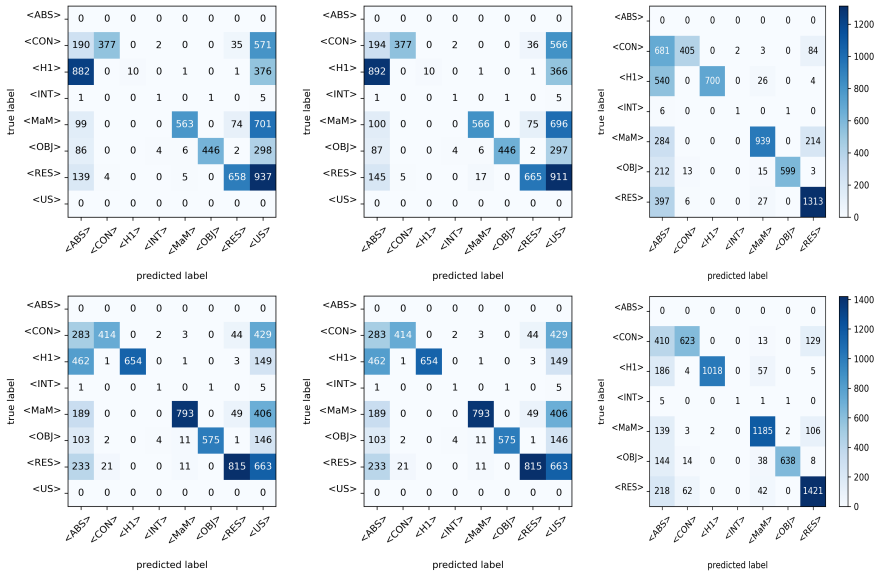

Figure 3: Confusion matrices for the Scielo testset (EN-FR): from left to right `pre0`, `pre1` and `pre2`. First row baseline, Second row fine-tuned systems

phrased in *Introduction* and *Conclusion* sections, whereas the illustrative and concrete diction exhibited by the words like *evidence* and *outcomes* have been mostly used in *{Conclusion, Results}* and *{Objectives, Material and Methods, Conclusion}* sections respectively. Interestingly the use of comparative words like *but* and *difference* was mostly in *Results* section. Words like *conclusion* and *controlled* have exhibited the highest probability of usage only in *Conclusion* and *Material and Methods* sections respectively. The confusion matrices and prediction scores show appreciable capability of the model for predicting the sections. We see that *Title* (<H1>) was predicted with the highest precision as titles are mostly content words. *Introduction* (<INT>) was the most difficult to predict as it typically presents the broader picture and has high lexical variation. *{Objectives, Material and Methods, Conclusion}* were predicted with precision above 88.

	EN-FR						FR-EN					
	Baseline			Fine-tuned			Baseline			Fine-tuned		
Section	P	R	F1	P	R	F1	P	R	F1	P	R	F1
<H1>	1.00	0.55	0.71	1.00	0.80	0.89	0.97	0.46	0.62	1.00	0.86	0.92
<OBJ>	1.00	0.71	0.83	0.99	1.00	0.76	1.00	0.57	0.72	0.99	0.74	0.85
<INT>	0.33	0.12	0.18	1.00	0.12	0.22	0.20	0.20	0.20	0.50	0.20	0.29
<MaM>	0.93	0.65	0.77	0.89	0.82	0.85	0.86	0.85	0.85	0.88	0.87	0.87
<CON>	0.96	0.34	0.51	0.88	0.53	0.66	0.99	0.32	0.49	0.88	0.53	0.67
<RES>	0.81	0.75	0.78	0.85	0.82	0.83	0.88	0.72	0.79	0.87	0.81	0.84
accuracy	-	-	0.61	-	-	0.75	-	-	0.61	-	-	0.77
macro avg	0.72	0.45	0.54	0.80	0.55	0.62	0.70	0.45	0.53	0.73	0.57	0.63
weighted avg	0.92	0.61	0.72	0.91	0.75	0.82	0.93	0.61	0.71	0.91	0.77	0.83

Table 7: Precision, Recall and F1 Score on Scielo baseline and fine-tuned systems for systems for section prediction (`pre2`).

When we predict all tags (`pre0`), systems have precision above 0.95 for *Conclusion, Title, Material and Methods, Objectives* and 0.85 for *Results*. In terms of accuracy, we see an improvement from 0.32 to 0.50 for EN-FR after fine-tuning and a slight improvement for FR-EN. As seen earlier, section prediction is better when the first two tags are known, i.e `pre2`, however title prediction remains good in FR-EN direction even for `pre0` and `pre1`. *Introduction* remains the most difficult section to predict but `pre2` systems learn to predict well. Accuracy for the baseline systems is 0.61, which increases for fine tuned systems to 0.75 and 0.77 for EN-FR and FR-EN respectively.

Word	Occurrence in Sections		
conclusion_NNS	CON		
controlled_JJ	MAM		
but_CC	RES		
difference_NN	RES		
however_RB	INT	CON	
may_MD	INT	CON	
trial_NNS	MAM	RES	
evidence_NN	RES	CON	
people_NNS	H1	OBJ	MAM
outcomes_NNS	OBJ	MAM	CON
studies_NNS	MAM	RES	CON
quality_NN	MAM	RES	CON

Table 8: Word occurrence in sections from 50 most frequent words only. POS tags were generated using nltk pos tagger.

6.3. Qualitative Analysis: Tag conditioned translation variants

In this section, we study how tags affect the lexical and grammatical choice in automatic translations. For this analysis, we analysed the 1167 translations of all the titles (<H1> tags) in the Scielo corpus and compared two cases: (a)

when the correct tag <H1> is used in source (b) when the generic tag (<ABS>) is used instead. Titles are interesting because they tend to be more compact in their wording than regular sentences and also differ in their syntactic structure as they mosly consist in long nominal phrases devoid of verbs (some illustrations in Table 9). As discussed above, titles can be predicted with some confidence (in pre2 conditions).

Overall, we observed that the two outputs were very comparable and hardly distinguishable in their BLEU scores (40.6 vs 40.9). In the majority of cases (931 titles), the two output translations are entirely identical. Out of the remaining cases, the differences are always small, yet we can identify interesting differences which illustrate the effect of these subsection tags. The first is that when using the <H1> tag, target titles are actually slightly shorter than in the other condition: the average length difference between the two outputs is 1.06 chars, with 77 cases of (a) being longer, and 131 cases where (b) is longer - these two situations are illustrated in the first two examples of Table 9.[17] There are various causes for these length differences, which are sometimes as subtle as a change in a determiner, see example 3 where translation (a) contains a collective use for the singular determiner, which is very appropriate in scientific texts; and example 4 where (a) uses a noun group while (b) uses a verbal form. One syntactic pattern however emerges from this analysis: outputs (a) tend to favour adjectives over prepositional groups (see examples 5 and 6).

More examples of variations obtained with different tags are in apprendix Appendix B.

7. Related Work

Document-level MT. (DLMT) is loosely defined by contrast to "sentence-level" MT (SLMT) and includes any technique aiming to handle translation phenomena requiring a context extending beyond isolated sentences. The targeted phenomena

[17] Indeed the BLEU scores between these condition is mostly due to a lower brevity penalty when using <H1>.

1 - src	Socioeconomic and geographic inequalities in infant mortality, 1990-2005
trg (a)	Inégalités socioéconomiques et géographiques de mortalité infantile, 1990-2005.
trg (b)	Inégalités socioéconomiques et géographiques **en matière** de mortalité infantile, 1990-2005.
2- src	The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000.
trg (a)	La charge **globale** des maladies diarrhéiques, **telle qu'estimée à partir d'**études publiées entre 1992 et 2000.
trg (b)	La charge **mondiale** des maladies diarrhéiques, **estimée d'après des** études publiées entre 1992 et 2000.
3 - src	Homicide in children and adolescents: a case-control study in Recife, Brazil.
trg (a)	Homicide chez **l'enfant et l'adolescent**: étude cas-témoins à Recife, Brésil.
trg (b)	Homicide chez **les enfants et les adolescents:** étude cas-témoins à Recife, Brésil.
4 - src	Managing the effect of TRIPS on availability of priority vaccines.
trg (a)	**Gestion de l'effet des ADPIC** sur la disponibilité des vaccins prioritaires.
trg (b)	**Gérer l'effet de l'ADPIC** sur la disponibilité des vaccins prioritaires.
5 - src	Poverty, child undernutrition and morbidity: new evidence from India.
trg (a)	Pauvreté, sous-nutrition infantile et morbidité: nouvelles données **indiennes**.
trg (b)	Pauvreté, sous-nutrition infantile et morbidité: nouveaux éléments **concernant l'Inde**.
6 - src	Circulating vaccine-derived polioviruses: current state of knowledge.
trg (a)	Poliovirus circulants **d'origine vaccinale**: état actuel des connaissances.
trg (b)	Poliovirus circulants **dérivés du vaccin:** état actuel des connaissances.

Table 9: Translation of Scielo test sets produced with fine-tuned systems from English into French. For each example, we display the source sentence, then the translations obtained respectively with the tags <H1> (a) and <ABS> (b). Note that for all these examples, the two French outputs are equally fluent and adequate. They illustrate small variations in wordings reflecting the style differences in titles vs. regular texts.

are quite heterogeneous in nature, but notably encompass coreference issues, coherence issues, and discourse-level issues [4]. Coreference issues are mostly related to the consistency of pronoun use across languages, where the choice of a morphological variant for a pronoun (e.g., in gender or number) may be conditioned by its referent from previous sentence(s). Coherence corresponds to longer range phenomena that ensure that a translated text can be read as a whole, implying a consistent choice of terms, tense, style, and references throughout the text. Finally, DLMT also includes the generation of texts that correctly reproduce the argumentative structure of the source input. In the realm of neural architectures, DLMT has been mostly approached as "MT with long-range dependencies", fostering multiple approaches to integrate large spans of text in the translation context. These approaches range from simple extensions of SLMT where multiple sentences are translated as one unit [60, 61, 62], to more sophisticated proposals combining a short-term context (at the sentence

level) with a long-range context. This can be achieved with dual encoders [63], hierarchical architectures [64, 65, 66], or cache-based methods [12]. Widening the notion of context, topic models or even domain adaptation techniques can also be used to represent large documents [9]. A conclusion of several comparative studies [67, 15, 62] is that simple techniques (no context, or basic concatenation) are difficult to outperform. A review of DLMT is in [6]; more recent approaches based on large language models are documented in [68].

Tags. have been widely used in NMT to incorporate additional discrete conditioning factors in (1), and control the output language, the domain, or the level of formality and politeness [29, 30, 31], mitigate gender bias issues [69], and provide us with a simple and effective way to take structure in account.

Notably, [30, 40] use tags in multi-domain systems, an approach extended in [70] with multi-level; tags in [71, 72, 73, 31] inform systems about the level of formality or politeness of a sentence; finally, multilingual MT [74, 29, 75] use tags to select the desired target language. In the same line of studies, [76] use numerical tag values corresponding to length constraints, while [40] use a tag to enforce the decoding direction of a bidirectional system, in [35, 77] tags inform back-translated data, [78] use tags to control the readability level of NMT output, [42, 79] demonstrate zero-shot NMT capability using tags.

Tags have also proven useful for grammatical error correction using NMT [80] where tags are used with each word to indicate the operation e.g. keep, delete, replace etc. [81] use tags to enforce the decoder towards natural MT output, [82] used tags for named entity recognition, whereas tags have been used in [83] for entity projection for cross-lingual NER.

To our knowledge, there exists no prior work on using tags to represent the document structure in NMT other than [77], which only use tags on the source side. Closer to our work, one-level document structure-tags have been used for document quality prediction by [84], which reports a strong correlation with text classification. They used three tags (Title, Abstract and BodyText) and reported performance reduction when using smaller structure-tag set on all three

domains that they worked on.

In the context of Multilingual NMT, [42] shows language tags to significantly impact zero-shot translation quality. They report best scores by placing the target language tags on the encoder side which helps alleviate the off-target issue for such models. Similar analyses, for a variety of tags, are in [40, 41], and suggest that when tags are observed, having them on the source side yields better results. Our results also show better results with tags on the source side.

It is further possible to explicitly amplify this effect by injecting the tag information into the input representation of every token (eg. [30, 36] for domain information, or [37] for language information).

8. Conclusion and Discussion

Sentences in a document generally follow a typical style, for example the introductory sentences are clearly distinguishable from concluding sentences based on the lexical choice and sentence style etc. Whether or not can an MT system be trained to learn and use these sentence specific styles is the question that we have studied in this article.

We have tried to take advantage of the rigid structure of abstracts in the biomedical domain to assess the effect of structure on the translation quality. For this, we have carefully annotated a large subset of our training data with structural information, and trained models that were able to take this structure into account thanks to a hierarchical tag system that introduces an extra non-local context during translation. In addition to the elaboration of a new annotated resource, that will be useful for further works, our study has shown that (a) predicting the structural labels was possible from the sole source text, at least for fine tune systems exposed to structured documents; (b) based on this complex of tags, it was actually possible to improve our automatic evaluation scores, even though the actual effect of the sole structural tags on BLEU scores was found to be fairly limited in our test conditions. This highlights one clear limitation of our results, which mostly rely on automatic metrics such as BLEU or TER. It is

likely that such scores may not fully capture the fine-grained stylistic variations that are implied by the use of document structure information.

In our future work, we would like to generalize this approach to other structural labels that may be available for other documents and and domains, as well as to combine this information with other, more fine grain, additional contextual information that may be useful for document-level MT. Another perspective is to explore how structural information is handled in NMT systems based on large language models, which represent a competitive alternative to the encoder-decoder models used in this study.

Acknowledgements

This work was made possible thanks to the Saclay-IA computing platform. It was granted access to the HPC resources of IDRIS under the allocation 2021-[AD011011580R1, AD011011270R1, AD011011717] made by GENCI. The second author was partly funded by the French "Agence Nationale de la Recherche" (ANR) under grant ANR-22-CE23-0033 / MaTOS.

References

[1] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, et al., Achieving human parity on automatic Chinese to English news translation, arXiv preprint arXiv:1803.05567 (2018).

[2] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, Z. Žabokrtský, Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals, Nature Communications 11 (1) (2020) 4381. doi:10.1038/s41467-020-18073-9. URL https://doi.org/10.1038/s41467-020-18073-9

[3] E. Voita, P. Serdyukov, R. Sennrich, I. Titov, Context-aware neural machine translation learns anaphora resolution, in: Proceedings of the 56th Annual

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1264–1274. doi:10.18653/v1/P18-1117.
URL https://www.aclweb.org/anthology/P18-1117

[4] A. Popescu-Belis, Context in neural machine translation: A review of models and evaluations, arXiv preprint arXiv:1901.09115 (2019).
URL http://arxiv.org/pdf/1901.09115

[5] R. Bawden, K. Bretonnel Cohen, C. Grozea, A. Jimeno Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, A. Siu, K. Verspoor, M. Vicente Navarro, Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies, in: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Association for Computational Linguistics, Florence, Italy, 2019, pp. 29–53. doi:10.18653/v1/W19-5403.
URL https://www.aclweb.org/anthology/W19-5403

[6] S. Maruf, F. Saleh, G. Haffari, A survey on document-level neural machine translation: Methods and evaluation, ACM Comput. Surv. 54 (2) (Mar. 2021). doi:10.1145/3441691.
URL https://doi.org/10.1145/3441691

[7] P. Fernandes, K. Yin, E. Liu, A. Martins, G. Neubig, When does translation require context? a data-driven, multilingual exploration, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 606–626.
URL https://aclanthology.org/2023.acl-long.36

[8] K. Hofmann, M. Tsagkias, E. Meij, M. De Rijke, The impact of document structure on keyphrase extraction, in: Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1725–1728.

[9] S. Kuang, D. Xiong, W. Luo, G. Zhou, Modeling coherence for neural machine translation with dynamic and topic caches, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 596–606.
URL https://www.aclweb.org/anthology/C18-1050

[10] J. Tiedemann, Y. Scherrer, Neural machine translation with extended context, in: Proceedings of the Third Workshop on Discourse in Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 82–92. doi:10.18653/v1/W17-4811.
URL https://www.aclweb.org/anthology/W17-4811

[11] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, Y. Liu, Improving the Transformer translation model with document-level context, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 533–542. doi:10.18653/v1/D18-1049.
URL https://www.aclweb.org/anthology/D18-1049

[12] Z. Tu, Y. Liu, S. Shi, T. Zhang, Learning to remember translation history with a continuous cache, Transactions of the Association for Computational Linguistics 6 (2018) 407–420. doi:10.1162/tacl_a_00029.
URL https://www.aclweb.org/anthology/Q18-1029

[13] M. Junczys-Dowmunt, Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Association for Computational Linguistics, Florence, Italy, 2019, pp. 225–233. doi:10.18653/v1/W19-5321.
URL https://www.aclweb.org/anthology/W19-5321

[14] Q. Guo, X. Qiu, P. Liu, Y. Shao, X. Xue, Z. Zhang, Star-transformer, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1315–1325. doi:10.18653/v1/N19-1133.
URL https://www.aclweb.org/anthology/N19-1133

[15] A. Lopes, M. A. Farajian, R. Bawden, M. Zhang, A. Martins, Document-level neural MT: A systematic comparison, in: 22nd Annual Conference of the European Association for Machine Translation, 2020, pp. 225–234.

[16] S. Ma, D. Zhang, M. Zhou, A simple and effective unified encoder for document-level machine translation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3505–3511. doi:10.18653/v1/2020.acl-main.321.
URL https://www.aclweb.org/anthology/2020.acl-main.321

[17] S. Abdul Rauf, F. Yvon, Document level contexts for neural machine translation, Research Report 2020-003, LIMSI-CNRS (Dec. 2020).
URL https://hal.archives-ouvertes.fr/hal-03687190

[18] K. Hashimoto, R. Buschiazzo, J. Bradbury, T. Marshall, R. Socher, C. Xiong, A high-quality multilingual dataset for structured documentation translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), Association for Computational Linguistics, Florence, Italy, 2019, pp. 116–127. doi:10.18653/v1/W19-5212.
URL https://aclanthology.org/W19-5212

[19] G. Hanneman, G. Dinu, How should markup tags be translated?, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 1160–1173.
URL https://aclanthology.org/2020.wmt-1.138

[20] R. Dobreva, J. Zhou, R. Bawden, Document sub-structure in neural machine translation, in: Proceedings of the 12th Language Resources and Evaluation

Conference, European Language Resources Association, Marseille, France, 2020, pp. 3657–3667.
URL https://www.aclweb.org/anthology/2020.lrec-1.451

[21] B. Haddow, A. Birch, K. Heafield, Machine translation in healthcare, in: The Routledge Handbook of Translation and Health, Routledge, 2021, pp. 108–129.

[22] M. Zappatore, G. Ruggieri, Adopting machine translation in the healthcare sector: A methodological multi-criteria review, Computer Speech & Language 84 (2024) 101582. doi:https://doi.org/10.1016/j.csl.2023.101582.
URL https://www.sciencedirect.com/science/article/pii/S0885230823001018

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

[24] P. Koehn, Neural Machine Translation, Cambridge University Press, 2020.

[25] F. Stahlberg, Neural machine translation: A review, Journal of Artificial Intelligence Review 69 (2020) 343–418.

[26] G. Wiher, C. Meister, R. Cotterell, On decoding strategies for neural text generators (2022). doi:10.48550/ARXIV.2203.15721.
URL https://arxiv.org/abs/2203.15721

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 5998–6008.
URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[28] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR abs/1409.0473 (2014).
URL http://arxiv.org/abs/1409.0473

[29] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation (2016). arXiv:1609.08144.
URL https://arxiv.org/abs/1609.08144

[30] C. Kobus, J. Crego, J. Senellart, Domain control for neural machine translation, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, 2017, pp. 372–378. doi:10.26615/978-954-452-049-6_049.
URL https://doi.org/10.26615/978-954-452-049-6_049

[31] A. Madaan, A. Setlur, T. Parekh, B. Poczos, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, S. Prabhumoye, Politeness transfer: A tag and generate approach, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1869–1881. doi:10.18653/v1/2020.acl-main.169.
URL https://aclanthology.org/2020.acl-main.169

[32] R. C. Moore, Fast and accurate sentence alignment of bilingual corpora, in: S. D. Richardson (Ed.), Proc. AMTA'02, Lecture Notes in Computer Science 2499, Springer Verlag, Tiburon, CA, USA, 2002, pp. 135–144.
URL https://www.microsoft.com/en-us/research/publication/fast-and-accurate-sentence-alignment-of-bilingual-corpora/

[33] W. Peng, J. Liu, L. Li, Q. Liu, 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), Association for Computational Linguistics, Florence, Italy, 2019. doi:10.18653/v1/

W19-5420.
URL https://www.aclweb.org/anthology/W19-5420

[34] I. Caswell, C. Chelba, D. Grangier, Tagged back-translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), Association for Computational Linguistics, Florence, Italy, 2019, pp. 53–63. doi:10.18653/v1/W19-5206.
URL https://aclanthology.org/W19-5206

[35] B. Marie, R. Rubino, A. Fujita, Tagged back-translation revisited: Why does it really work?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5990–5997. doi:10.18653/v1/2020.acl-main.532.
URL https://aclanthology.org/2020.acl-main.532

[36] C. Chu, R. Dabre, S. Kurohashi, An empirical comparison of domain adaptation methods for neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2017, Vancouver, Canada, 2017, pp. 385–391.
URL http://aclweb.org/anthology/P17-2061

[37] A. Conneau, G. Lample, Cross-lingual language model pretraining, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 7059–7069.
URL http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf

[38] D. Britz, Q. Le, R. Pryzant, Effective domain mixing for neural machine translation, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 118–126.
URL http://aclweb.org/anthology/W17-4712

[39] J. Wuebker, S. Green, J. DeNero, S. Hasan, M.-T. Luong, Models and inference for prefix-constrained machine translation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 66–75. doi:10.18653/v1/P16-1007.
URL https://aclanthology.org/P16-1007

[40] S. Takeno, M. Nagata, K. Yamamoto, Controlling target features in neural machine translation via prefix constraints, in: Proceedings of the 4th Workshop on Asian Translation (WAT2017), Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 55–63.
URL https://aclanthology.org/W17-5702

[41] M. Q. Pham, J. Crego, F. Yvon, Revisiting multi-domain machine translation, Transactions of the Association for Computational Linguistics 9 (0) (2021) 17–35.
URL https://transacl.org/index.php/tacl/article/view/2327

[42] L. Wu, S. Cheng, M. Wang, L. Li, Language tags matter for zero-shot neural machine translation, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3001–3007. doi:10.18653/v1/2021.findings-acl.264.
URL https://aclanthology.org/2021.findings-acl.264

[43] A. Jimeno Yepes, A. Névéol, M. Neves, K. Verspoor, O. Bojar, A. Boyer, C. Grozea, B. Haddow, M. Kittner, Y. Lichtblau, P. Pecina, R. Roller, R. Rosa, A. Siu, P. Thomas, S. Trescher, Findings of the WMT 2017 biomedical translation shared task, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 234–247. doi:10.18653/v1/W17-4719.
URL https://www.aclweb.org/anthology/W17-4719

[44] M. Neves, A. J. Yepes, A. Névéol, 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), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 2942–2948.
URL https://www.aclweb.org/anthology/L16-1470

[45] J. Ive, A. Max, F. Yvon, P. Ravaud, 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), Portorož, Slovenia, 2016, p. 8.
URL http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MT%20Evaluation_Proceedings.pdf#page=65

[46] P. Lison, J. Tiedemann, OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 923–929.
URL https://www.aclweb.org/anthology/L16-1147

[47] S. Abdul-Rauf, H. Schwenk, 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), Association for Computational Linguistics, Athens, Greece, 2009, pp. 16–23.
URL https://www.aclweb.org/anthology/E09-1003

[48] S. Naz, S. Abdul Rauf, N.-e. Hira, S. Ul Haq, FJWU participation for the WMT20 biomedical translation task, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 849–856.
URL https://aclanthology.org/2020.wmt-1.92

[49] M. Neves, A. Jimeno Yepes, A. Névéol, C. Grozea, A. Siu, M. Kittner, K. Verspoor, Findings of the WMT 2018 biomedical translation

shared task: Evaluation on Medline test sets, in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 324–339. doi:10.18653/v1/W18-6403.
URL https://aclanthology.org/W18-6403

[50] R. Bawden, K. Bretonnel Cohen, C. Grozea, A. Jimeno Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, A. Siu, K. Verspoor, M. Vicente Navarro, Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies, in: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Association for Computational Linguistics, Florence, Italy, 2019, pp. 29–53. doi:10.18653/v1/W19-5403.
URL https://aclanthology.org/W19-5403

[51] D. Stojanovski, V. Hangya, M. Huck, A. Fraser, 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), Association for Computational Linguistics, Florence, Italy, 2019, pp. 393–399. doi:10.18653/v1/W19-5344.
URL https://www.aclweb.org/anthology/W19-5344

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

[53] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 48–53. doi:10.18653/v1/N19-4009.
URL https://www.aclweb.org/anthology/N19-4009

[54] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
URL http://arxiv.org/abs/1412.6980

[55] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. doi:10.18653/v1/D18-2012.
URL https://www.aclweb.org/anthology/D18-2012

[56] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 186–191. doi:10.18653/v1/W18-6319.
URL https://www.aclweb.org/anthology/W18-6319

[57] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Stroudsburg, PA, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135.
URL https://www.aclweb.org/anthology/P02-1040

[58] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA), Boston, Massachusetts, USA, 2006, pp. 223–231.
URL http://www.mt-archive.info/AMTA-2006-Snover.pdf

[59] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation,

Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. doi:10.18653/v1/W15-3049.
URL https://aclanthology.org/W15-3049

[60] Y. Scherrer, J. Tiedemann, S. Loáiciga, Analysing concatenation approaches to document-level NMT in two different domains, in: Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 51–61. doi:10.18653/v1/D19-6506.
URL https://www.aclweb.org/anthology/D19-6506

[61] Z. Ma, S. Edunov, M. Auli, A comparison of approaches to document-level machine translation, arXiv preprint arXiv:1910.07481 (2021). arXiv:2101.11040.

[62] Z. Sun, M. Wang, H. Zhou, C. Zhao, S. Huang, J. Chen, L. Li, Rethinking document-level neural machine translation, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3537–3548. doi:10.18653/v1/2022.findings-acl.279.
URL https://aclanthology.org/2022.findings-acl.279

[63] R. Bawden, R. Sennrich, A. Birch, B. Haddow, Evaluating discourse phenomena in neural machine translation, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1304–1313. doi:10.18653/v1/N18-1118.
URL https://www.aclweb.org/anthology/N18-1118

[64] L. Miculicich, D. Ram, N. Pappas, J. Henderson, Document-level neural machine translation with hierarchical attention networks, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp.

2947–2954. doi:10.18653/v1/D18-1325.
URL https://www.aclweb.org/anthology/D18-1325

[65] J. Chen, X. Li, J. Zhang, C. Zhou, J. Cui, B. Wang, J. Su, Modeling discourse structure for document-level neural machine translation, in: Proceedings of the First Workshop on Automatic Simultaneous Translation, Association for Computational Linguistics, Seattle, Washington, 2020, pp. 30–36. doi:10.18653/v1/2020.autosimtrans-1.5.
URL https://www.aclweb.org/anthology/2020.autosimtrans-1.5

[66] Z. Zheng, X. Yue, S. Huang, J. Chen, A. Birch, Towards making the most of context in neural machine translation, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 3983–3989. doi:10.24963/ijcai.2020/551.
URL https://doi.org/10.24963/ijcai.2020/551

[67] Y. Kim, D. T. Tran, H. Ney, When and why is document-level context useful in neural machine translation?, in: Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 24–34. doi:10.18653/v1/D19-6503.
URL https://www.aclweb.org/anthology/D19-6503

[68] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, Z. Tu, Document-level machine translation with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 16646–16661. doi:10.18653/v1/2023.emnlp-main.1036.
URL https://aclanthology.org/2023.emnlp-main.1036

[69] D. Saunders, R. Sallis, B. Byrne, Neural machine translation doesn't translate gender coreference right unless you make it (2020). arXiv:2010.05332.

[70] E. Stergiadis, S. Kumar, F. Kovalev, P. Levin, Multi-domain adaptation in neural machine translation through multidimensional tagging, CoRR abs/2102.10160 (2021). arXiv:2102.10160.
URL https://arxiv.org/abs/2102.10160

[71] R. Sennrich, B. Haddow, Linguistic input features improve neural machine translation, in: Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 83–91. doi:10.18653/v1/W16-2209.
URL https://www.aclweb.org/anthology/W16-2209

[72] X. Niu, M. Martindale, M. Carpuat, A study of style in machine translation: Controlling the formality of machine translation output, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2814–2819. doi:10.18653/v1/D17-1299.
URL https://www.aclweb.org/anthology/D17-1299

[73] X. Niu, S. Rao, M. Carpuat, Multi-task neural models for translating between styles within and across languages, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1008–1021.
URL https://aclanthology.org/C18-1086

[74] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 866–875. doi:10.18653/v1/N16-1101.
URL http://www.aclweb.org/anthology/N16-1101

[75] M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. a. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, Google's

multilingual neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351.
URL https://transacl.org/ojs/index.php/tacl/article/view/1081

[76] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, M. Okumura, Controlling output length in neural encoder-decoders, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 1328–1338. doi:10.18653/v1/D16-1140.
URL http://aclweb.org/anthology/D16-1140

[77] J. Xu, M. Q. Pham, S. Abdul Rauf, F. Yvon, LISN @ WMT 2021, in: Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, Online, 2021, pp. 232–242.
URL https://aclanthology.org/2021.wmt-1.22

[78] K. Marchisio, J. Guo, C.-I. Lai, P. Koehn, Controlling the reading level of machine translation output, in: Proceedings of Machine Translation Summit XVII Volume 1: Research Track, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 193–2004.
URL https://www.aclweb.org/anthology/W19-6619

[79] Z. Mao, R. Dabre, Q. Liu, H. Song, C. Chu, S. Kurohashi, Exploring the impact of layer normalization for zero-shot neural machine translation (2023). arXiv:2305.09312.

[80] D. Liang, C. Zheng, L. Guo, X. Cui, X. Xiong, H. Rong, J. Dong, BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis, in: Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, Association for Computational Linguistics, Suzhou, China, 2020, pp. 57–66.
URL https://aclanthology.org/2020.nlptea-1.8

[81] M. Freitag, D. Vilar, D. Grangier, C. Cherry, G. Foster, A natural diet: Towards improving naturalness of machine translation output, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3340–3353. doi:10.18653/v1/2022.findings-acl.263.
URL https://aclanthology.org/2022.findings-acl.263

[82] A. Jain, B. Paranjape, Z. C. Lipton, Entity projection via machine translation for cross-lingual NER, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1083–1092. doi:10.18653/v1/D19-1100.
URL https://aclanthology.org/D19-1100

[83] A. Berard, I. Calapodescu, C. Roux, Naver labs Europe's systems for the WMT19 machine translation robustness task, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Association for Computational Linguistics, Florence, Italy, 2019, pp. 526–532. doi:10.18653/v1/W19-5361.
URL https://aclanthology.org/W19-5361

[84] G. Maillette de Buy Wenniger, T. van Dongen, E. Aedmaa, H. T. Kruitbosch, E. A. Valentijn, L. Schomaker, Structure-tags improve text classification for scholarly document quality prediction, in: Proceedings of the First Workshop on Scholarly Document Processing, Association for Computational Linguistics, Online, 2020, pp. 158–167. doi:10.18653/v1/2020.sdp-1.18.
URL https://aclanthology.org/2020.sdp-1.18

Appendix A. Precision and Recall scores for Confusion matrices

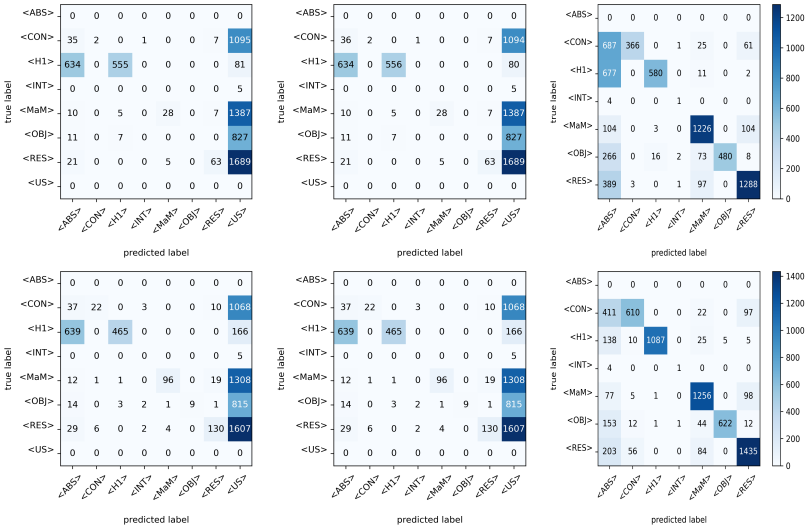

Figure A.4: FR-EN: from left to right `pre0`, `pre1` and `pre2`. First row baseline, Second row fine-tuned systems

Prefix	Section	EN-FR						FR-EN					
		Baseline			Fine-tuned			Baseline			Fine-tuned		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pre0	<CON>	0.99	0.32	0.48	0.95	0.35	0.51	1.00	0.00	0.00	0.76	0.02	0.04
	<H1>	1.00	0.01	0.02	1.00	0.51	0.68	0.98	0.44	0.60	0.99	0.37	0.53
	<INT>	0.14	0.12	0.13	0.14	0.12	0.13	0.00	0.00	0.00	0.00	0.00	0.00
	<MaM>	0.98	0.39	0.56	0.97	0.55	0.70	0.85	0.02	0.04	0.95	0.07	0.12
	<OBJ>	1.00	0.53	0.69	1.00	0.68	0.81	0.00	0.00	0.00	1.00	0.01	0.02
	<RES>	0.85	0.38	0.52	0.89	0.47	0.61	0.82	0.04	0.07	0.81	0.07	0.13
	accuracy	-	-	0.32	-	-	0.50	-	-	0.10	-	-	0.11
	macro avg	0.62	0.22	0.30	0.62	0.34	0.43	0.46	0.06	0.09	0.56	0.07	0.11
	weighted avg	0.95	0.32	0.45	0.95	0.50	0.65	0.78	0.10	0.15	0.89	0.11	0.18
Pre1	<CON>	0.99	0.32	0.48	0.95	0.35	0.51	1.00	0.00	0.00	0.76	0.02	0.04
	<H1>	1.00	0.01	0.02	1.00	0.51	0.68	0.98	0.44	0.61	0.99	0.37	0.53
	<INT>	0.14	0.13	0.08	0.14	0.12	0.13	0.00	0.00	0.00	0.00	0.00	0.00
	<MaM>	0.96	0.39	0.56	0.97	0.55	0.70	0.85	0.02	0.04	0.95	0.07	0.12
	<OBJ>	1.00	0.53	0.69	1.00	0.68	0.81	0.00	0.00	0.00	1.00	0.01	0.02
	<RES>	0.85	0.38	0.53	0.89	0.47	0.61	0.82	0.04	0.07	0.81	0.07	0.13
	accuracy	-	-	0.32	-	-	0.50	-	-	0.10	-	-	0.11
	macro avg	0.62	0.22	0.30	0.62	0.34	0.43	0.45	0.06	0.09	0.56	0.07	0.11
	weighted avg	0.95	0.32	0.45	0.95	0.50	0.65	0.78	0.10	0.13	0.89	0.11	0.18

Table A.10: Precision, Recall and F1 Score on Scielo baseline and fine-tuned systems without abstract in reference

Prefix	Section	EN-FR						FR-EN					
		Baseline			Fine-tuned			Baseline			Fine-tuned		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pre2	<CON>	0.96	0.34	0.51	0.88	0.53	0.66	0.99	0.32	0.49	0.88	0.53	0.67
	<H1>	1.00	0.55	0.71	1.00	0.80	0.89	0.97	0.46	0.62	1.00	0.86	0.92
	<INT>	0.33	0.12	0.18	1.00	0.12	0.22	0.20	0.20	0.20	0.50	0.20	0.29
	<MaM>	0.93	0.65	0.77	0.89	0.82	0.85	0.86	0.85	0.85	0.88	0.87	0.87
	<OBJ>	1.00	0.71	0.83	0.99	1.00	0.76	1.00	0.57	0.72	0.99	0.74	0.85
	<RES>	0.81	0.75	0.78	0.85	0.82	0.83	0.88	0.72	0.79	0.87	0.81	0.84
	accuracy	-	-	0.61	-	-	0.75	-	-	0.61	-	-	0.77
	macro avg	0.72	0.45	0.54	0.80	0.55	0.62	0.70	0.45	0.53	0.73	0.57	0.63
	weighted avg	0.92	0.61	0.72	0.91	0.75	0.82	0.93	0.61	0.71	0.91	0.77	0.83
Pre2 with <ABS>	<ABS>	0.48	0.82	0.61	0.62	0.78	0.70	0.48	0.85	0.61	0.64	0.76	0.70
	<CON>	0.85	0.40	0.54	0.68	0.53	0.60	0.97	0.32	0.48	0.70	0.54	0.61
	<H1>	1.00	0.09	0.17	1.00	0.80	0.89	0.96	0.46	0.62	0.99	0.86	0.92
	<INT>	0.01	0.12	0.02	0.25	0.12	0.17	0.07	0.20	0.11	0.50	0.20	0.29
	<MaM>	0.74	0.80	0.77	0.81	0.82	0.82	0.76	0.85	0.80	0.78	0.87	0.83
	<OBJ>	0.99	0.71	0.83	0.99	0.76	0.86	1.00	0.57	0.72	0.99	0.74	0.84
	<RES>	0.74	0.82	0.78	0.77	0.82	0.79	0.80	0.72	0.76	0.77	0.81	0.79
	accuracy	-	-	0.64	-	-	0.76	-	-	0.67	-	-	0.77
	macro avg	0.69	0.54	0.53	0.73	0.66	0.69	0.72	0.57	0.59	0.77	0.68	0.71
	weighted avg	0.75	0.64	0.61	0.78	0.76	0.76	0.77	0.67	0.67	0.78	0.77	0.77

Table A.11: Precision, Recall and F1 Score on Scielo baseline and fine-tuned systems for systems with and without abstract in reference comparison

Appendix B. Examples of variations incurred by tag differences

For this analysis, we generated translations of random documents using all possible tag combinations in source and target sides. A large set of these combinations were human analysed, some sample outputs for three tag combinations are in Table B.12 for SRCTAG vs TRGTAG with the corresponding BLEU, TER and chrF scores.

We observe variation in lexical choice as well as grammatical sentence formation, due to using SRCTAG vs TRGTAG. We see the preference of all SRCTAG combinations for certain lexical choices, e.g. SRCTAG system outputs always starts with the article `the`. Similarly, several phrase constructions appear to have been learned from out-of-domain data as these majorly appear in SRCTAG sentences having *any* tag from {<G> <WMT> <US> }.[18] For instance, "death registration" and "has been" in second and third examples.

[18]This tag combination is used for all out-of-domain sentences in the training corpus.

Source: La formation d'agents sanitaires de proximité, de médecins et de codeurs s'est révélée pour améliorer l'enregistrement des décès et accroître la disponibilité de données sur les causes de décès. **Reference:** Training community-based health workers, physicians and coders were successful approaches to improve death registration completeness and availability of cause-of-death data.
<M> <MED> <ABS> **SRCTAC:** the training of community health workers, physicians and coders has proven to be effective in improving the registration of deaths and increasing the availability of data on causes of death. **{BLEU:36.9 TER: 50.1 chrF:66.3}** **TRCTAC reference:** Training of outreach health workers, physicians and coders proved to be effective in improving the registration of deaths and increasing the availability of data on causes of death. **{BLEU:38.2 TER: 50.0 chrF:67.0}** **TRCTAC predicted:** Training of nearby health workers, doctors and coders has proven to be effective in improving the registration of deaths and increasing the availability of data on causes of death. **{BLEU:36.9 TER: 50.4 chrF:66.5}**
<M> <WMT> <CON> **SRCTAC:** The training of community health workers, physicians and coders has been effective in improving death registration and increasing the availability of data on causes of death. **{BLEU:38.2 TER: 48.7 chrF:66.9}** **TRCTAC reference:** Training of nearby health workers, physicians and coders has proven to be effective in improving the registration of deaths and increasing the availability of data on causes of death. **{BLEU:37.4 TER: 51.1 chrF:66.7}** **TRCTAC predicted:** Training of nearby health workers, doctors and coders has proven to be effective in improving the registration of deaths and increasing the availability of data on causes of death.**{BLEU:36.9 TER: 50.4 chrF:66.5}**
<G> <PATR> <ABS> **SRCTAC:** the formation of nearby health workers, physicians and encoders has been effective in improving death registration and increasing the availability of data on causes of death. **TRCTAC reference :** Training of outreach health workers, physicians and coders has proven to be effective in improving the registration of deaths and increasing the availability of data on causes of death. **TRCTAC predicted:** Training of nearby health workers, doctors and coders has proven to be effective in improving the registration of deaths and increasing the availability of data on causes of death.

Table B.12: Example outputs with varying tag combinations. Lexical differences are marked in blue whereas grammatical changes are given in orange. Invalid tags in a particular combination are marked in red. BLEU, TER and chrF scores are computed for whole test-set.

An interesting example is the <G> <PATR>[19] <ABS> combination (last row in TableB.12), where two tags are wrong and the SRCTAG system used the word "formation", "nearby" and "encoders" instead respectively of "training", "outreach" and "coders" which are generated with the reference tags.

[19]PATR is a subpart of the Ufal corpus amounting to around 22% of bio-medical data without substructure information, was thus tagged with <M> <PATR> <US>.

Appendix C. Number Similar and Different sentences selected in each decoding setting

Direction	Models	Testsets	Pre0-Pre1		Pre0-Pre2		Pre1-Pre2		Pre0-Pre3		Pre1-Pre3		Pre2-Pre3	
			Similar	Different	Similar	Different	Similar	Different	Similar	Different	Similar	Different	Similar	Different
en-fr	tagall	medline20	99.16%	0.84%	60.00%	40.00%	60.00%	40.00%	30.80%	69.20%	30.80%	69.20%	37.87%	62.13%
			(39.2)	(45.2 - 45.3)	(44.9)	(38.3 - 39.6)	(44.9)	(38.3 - 39.5)	(46.6)	(38.8 - 39.6)	(46.6)	(38.8 - 39.6)	(46.3)	(39.4 - 39.1)
		scielo	98.67%	1.33%	56.93%	43.07%	57.29%	42.71%	32.98%	67.02%	33.20%	66.80%	36.04%	63.96%
			(36.0)	(32.4 - 33.2)	(38.7)	(34.2 - 35.1)	(38.8)	(34.2 - 35.1)	(43.2)	(35.0 - 36.2)	(43.2)	(35.0 - 36.1)	(41.7)	(35.6 - 36.1)
		edp17	98.71%	1.29%	40.03%	59.97%	40.19%	59.81%	31.27%	68.73%	31.43%	68.57%	42.88%	57.12%
			(31.4)	(29.3 - 30.6)	(38.0)	(30.3 - 33.7)	(38.0)	(30.3 - 33.7)	(39.5)	(30.9 - 34.2)	(39.9)	(30.9 - 34.2)	(37.6)	(33.5 - 33.9)
	tagall-ftbiomed	medline20	100%	0%	53.69%	46.31%	53.69%	46.31%	50.04%	49.96%	50.04%	49.96%	84.64%	15.36%
			(40.2)	-	(43.1)	(38.7 - 40.3)	(43.1)	(38.7 - 40.3)	(44.4)	(38.4 - 39.4)	(44.4)	(38.4 - 39.4)	(42.4)	(37.3 - 35.6)
		scielo	100%	0%	53.40%	46.60%	53.40%	46.60%	52.54%	47.46%	52.54%	47.46%	8.86%	91.14%
			(35.5)	-	(38.4)	(34.1 - 36.0)	(38.4)	(34.1 - 36.0)	(38.2)	(34.2 - 36.0)	(38.2)	(34.2 - 36.0)	(37.1)	(34.4 - 33.8)
		edp17	100%	0%	43.95%	56.05%	43.95%	56.05%	43.58%	56.42%	43.58%	56.42%	95.00%	5.00%
			(34.4)	-	(42.1)	(32.7 - 36.8)	(42.1)	(32.7 - 36.8)	(41.8)	(32.8 - 36.8)	(41.8)	(32.8 - 36.8)	(37.9)	(37.2 - 35.7)
fr-en	tagall	medline20	99.85%	0.15%	64.41%	35.59%	64.33%	35.67%	58.48%	41.52%	58.40%	41.60%	83.50%	16.60%
			(40.0)	(43.3 - 47.4)	(40.9)	(38.1 - 39.2)	(41.0)	(39.2 - 39.2)	(41.4)	(38.9 - 39.1)	(41.5)	(38.9 - 39.1)	(41.0)	(35.9 - 36.2)
		scielo	99.86%	0.14%	52.77%	47.23%	52.78%	47.22%	52.27%	47.73%	52.28%	47.72%	92.33%	7.67%
			(35.2)	(44.2 - 43.1)	(38.9)	(33.5 - 35.0)	(38.9)	(33.5 - 35.0)	(39.0)	(33.5 - 35.0)	(39.0)	(33.5 - 35.0)	(36.7)	(32.6 - 32.7)
		edp17	99.57%	0.43%	59.38%	40.62%	59.43%	40.57%	58.46%	41.54%	58.41%	41.59%	92.37%	7.63%
			(32.2)	(41.6 - 42.1)	(34.6)	(30.7 - 31.0)	(34.6)	(30.7 - 31.0)	(34.7)	(30.7 - 31.0)	(34.6)	(30.7 - 31.1)	(32.9)	(28.4 - 28.7)
	tagall-ftbiomed	medline20	100%	0%	66.31%	33.69%	66.31%	33.69%	60.23%	39.77%	60.23%	39.77%	84.41%	15.59%
			(42.7)	-	(43.6)	(41.6 - 42.7)	(43.6)	(41.6 - 42.7)	(44.4)	(41.0 -42.0)	(44.4)	(41.0 -42.0)	(44.3)	(37.7 - 37.7)
		scielo	100%	0%	46.76%	53.24%	46.76%	53.24%	46.56%	53.44%	46.56%	53.44%	91.15%	8.85%
			(36.0)	-	(41.0)	(34.4 - 36.7)	(41.0)	(34.4 - 36.7)	(41.3)	(34.4 - 36.7)	(41.3)	(34.4 - 36.7)	(38.4)	(33.2 - 33.7)
		edp17	100%	0%	55.83%	44.17%	55.83%	44.17%	56.10%	43.90%	56.10%	43.90%	90.22%	9.78%
			(34.1)	-	(38.4)	(31.7 - 33.6)	(38.4)	(31.7 - 33.6)	(38.6)	(31.5 - 33.2)	(38.6)	(31.5 - 33.2)	(36.0)	(29.8 - 28.7)

Table C.13: Sentence Comparison, Percentage and BLEU scores