IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 10, 2021, accepted February 26, 2021, date of publication March 4, 2021, date of current version March 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063715

Machine Translation of Mathematical Text

ADITYA OHRI AND TANYA SCHMAH
Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada
Corresponding author: Tanya Schmah (tschmah@uottawa.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2016-06742.

ABSTRACT We have implemented a machine translation system, the PolyMath Translator, for LaTeX documents containing mathematical text. The current implementation translates English LaTeX to French LaTeX, attaining a BLEU score of 53.6 on a held-out test corpus of mathematical sentences. It produces LaTeX documents that can be compiled to PDF without further editing. The system first converts the body of an input LaTeX document into English sentences containing math tokens, using the pandoc universal document converter to parse LaTeX input. We have trained a Transformer-based translator model, using OpenNMT, on a combined corpus containing a small proportion of domain-specific sentences. Our full system uses this Transformer model and also Google Translate with a custom glossary, the latter being used as a backup to better handle linguistic features that do not appear in our training dataset. Google Translate is used when the Transformer model does not have confidence in its translation, as determined by a high perplexity score. Ablation testing demonstrates that the tokenization of symbolic expressions is essential to the high quality of translations produced by our system. We have published our test corpus of mathematical text. The PolyMath Translator is available as a web service at www.polymathtrans.ai.

INDEX TERMS Machine translation, natural language processing, multi-layer neural network, LaTeX.

I. INTRODUCTION

Machine translation for specialized domains such as legal or medical text has received considerable attention. Advances in these areas have been useful in practice and have also given rise to new techniques in areas including domain adaptation [4], automatic term extraction [25] and domain-aware approaches to general-purpose machine translation [2]. In the present paper we consider the domain of mathematical text, as produced by researchers in mathematics and related fields, and post-secondary teachers of these subjects. To clarify, we consider specialized natural language about mathematics, which we may call informal mathematical writing (even if highly technical), to distinguish it from formal mathematics written in the language of pure logic. In particular, the present paper considers the problem of translation of mathematical writing from English to French. The specific choice of language pair is motivated by the authors' particular Canadian context. The domain of mathematical text has, to our knowledge, not yet been the subject of research in machine translation, beyond some very early work [14], [17], although we mention extensive ongoing research in the related areas of mathematical ontology and semantics [31], translation from informal mathematical writing into formal mathematics [29], and mathematical information retrieval [10].

Mathematical text presents several features of interest to the researcher. Most obviously, it often mixes natural language with symbolic expressions (i.e. formulae) in the same sentence, with the symbolic expressions playing a variety of grammatical roles, including as nouns, pronouns or clauses. Mathematical text has its own grammar and conventions, e.g. "Let x and y be integers" or "Consider the function h = f+g." (See also [23].) We hypothesize that most symbolic expressions in mathematical text functions grammatically as nouns, and thus may be adequately replaced by single-word tokens for the purpose of machine translation of the surrounding language. Tokenization of entire symbolic expressions has been used by other researchers, for example in the production of the OPUS "Wikipedia" corpus [30], which we use in the present paper. There are of course many examples of symbolic expressions that do *not* function grammatically as nouns. For example in the sentence "Since x>y, it follows that x is positive.", the first symbolic expression "x>y" is a clause. For this and other reasons, we expect that grammatical and semantic understanding of symbolic expressions will eventually improve machine translation of mathematical text. However this is beyond the scope of the present article, which relies on simple tokenization of whole symbolic expressions.

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

Anecdotally, mathematical writing can be both more and less complex than typical natural language. It is unique in its frequent use of definitions [31], and often complex in its precise description of logical relationships, using vocabulary and grammar that is standard but uncommon outside of mathematics, philosophy and law. While we are not aware of research directly supporting this assertion, it is indirectly supported by Yasseri *et al.*'s study of Wikipedia entries [32], which found that the most complex examples studied, as measured by the Gunning fog index, were those in Philosophy and Physics. (Mathematics was not considered.) Note that the Gunning fog index depends on mean sentence length and the frequency of long words. On the other hand, there is no suggestion in that paper that the *vocabulary* of mathematical text is larger than usual. On the contrary, anecdotal evidence suggests the opposite. In particular, the authors are aware of many examples of mathematical researchers giving comprehensible technical lectures in languages in which they are not generally fluent. If this simplicity of vocabulary is confirmed, then the domain of mathematical text offers an intriguing possibility of better-than-usual machine translation in this limited domain, while at the same time possessing unique challenges.

Practically, there is a great need for machine translation of mathematical text, for researchers, teachers and students. While the commercial possibilities of this domain may not be as obvious as in medicine or law, there is a large base of potential users, and a very large number of documents. For example the MathSciNet® database [24] contains over 3.6 million items, and if we expand the definition of mathematical text to include mathematically rich fields such as physics, computer science and engineering, we may consider the 1.8 million papers on the `arXiv.org` preprint server. In addition, there are millions of students of advanced mathematics with limited fluency in English who rely on English-language textbooks because of the lack of available translations.

The most salient obstacle to machine translation of mathematical text has been its symbolic content. The vast majority of mathematical documents are now written in LaTeX, a document-preparation system specific to mathematics, which supports embedded mathematical symbols and expressions, as well as providing document layout features. This ubiquity is at once an obstacle to translation via existing tools (which do not understand LaTeX syntax) and an opportunity, since once the LaTeX "hurdle" is passed, the majority of modern mathematical text is now available to machine translation.

In this paper, we describe and evaluate a system for translating mathematical text, which includes LaTeX parsing using pandoc [16], math tokenization, and sentence translation using the "Transformer" neural machine translation model [27]. To our knowledge, this is the first system of its kind in the literature. We demonstrate that our system produces high-quality translations and that math tokenization is essential to achieving this. We also examine how mathematical text differs from text in other domains, and provide evidence that mathematical text is simpler in some respects. We publish a small corpus of English-French sentence pairs with math tokenized.

The rest of the paper is organized as follows. In the Methods section we outline the custom corpora and glossary that we use for model training and testing, and we then describe our algorithms. The Results section begins with a brief comparison of corpora in three specialized domains: mathematics, geography and sociology. We then evaluate our translation system on both whole LaTeX documents and a small test corpus of mathematical sentence pairs, and perform ablation testing to evaluate the importance of different components of our system. The paper ends with a Discussion.

II. METHODS

A. DATA AND PREPROCESSING

We constructed a small custom glossary of 373 mathematical terms (words and short phrases) by combining several publicly available lists (including [6], [22]) and adding a few extra terms.

Since we are not aware of any published corpora specifically designed for evaluating mathematical translation, we prepared three corpora of aligned English and French sentences: (i) "math-wiki" – a subset of the OPUS "Wikipedia" corpus (v1.0) [30]; (ii) "custom math", based on the second author's research papers in mathematics; (iii) "linear code", based on course notes by M. Nevins on linear error-correcting codes. All of these are described in detail below. We have published the "linear code" corpus on the IEEE DataPort [18].

The OPUS "Wikipedia" corpus is a corpus of parallel sentences extracted from Wikipedia by Krzysztof Wołk and Krzysztof Marasek [30], which is part of the OPUS project [26]. The full corpus includes many language pairs, of which we use only the English/French pair, which consists of 803,670 sentence pairs containing 34M words (English plus French). The subject matter is wide-ranging. In order to focus on mathematical text, we applied a naïve subject matter filter to this corpus. We extracted from the "Wikipedia" corpus only those sentence pairs in which the English sentence contained at least two terms from our custom mathematical glossary. We found 16,767 sentence pairs satisfying this criterion. We call the resulting sub-corpus "math-wiki". All symbolic expressions in this corpus, as in the original "Wikipedia" corpus, are tokenized. Due to the method used to select sentences, some non-mathematical sentences are included. The vocabulary of this reduced dataset contains 55,474 unique tokens, whereas the mean vocabulary size of random subsets of the same corpus, of the same size, was 100,738 (mean over 5 random samples).

For comparison, we also extracted subsets of the "Wikipedia" corpus relevant to the domains of geography and sociology, using the method described above, based on word lists extracted from: the Wikipedia entry "Glossary of geography terms", and a glossary by the American Sociological Society [1]. As for the "math-wiki" corpus, we noticed that our "geography" and "sociology" sub-corpora contain

some sentences outside the target domain, so are most accurately described as multi-domain corpora heavily weighted towards the target subject.

For all model training and validation, and for most of our testing, we used a combined corpus of pairs of aligned text chunks (mostly sentences), called hereafter our ''main corpus'', consisting of: the ''math-wiki'' corpus defined above; a subset of the ''Aligned Hansards of the 36th Parliament of Canada'' corpus [8]; and our own ''custom math'' corpus derived mainly from several research papers of the second author TS. The first sub-corpus (Hansards) was included to provide greater breadth of vocabulary, grammar and style, while the second two focus specifically on mathematical text.

The ''Aligned Hansards of the 36th Parliament of Canada'' corpus is a corpus of aligned text chunks (sentences or smaller fragments) extracted from the official records of the 36th Canadian Parliament, including debates from the House of Commons and the Senate [8]. The full corpus consists of 1.28M English-French sentence pairs, containing 33.9M words (English plus French). This is a high-quality corpus consisting mostly of complete sentences. We chose it for its size and quality, and also in the hope that the source material would contain many examples of a formal expository style of language with a structure similar to mathematical text. After removing sentence pairs with irrelevant information such as the title, date, and speaker names, we randomly shuffled the entire filtered corpus and then selected 250,000 sentence pairs.

The custom math corpus consists of 1,075 sentence pairs in English and French, with the English sentences extracted from several of the second author's research papers in mathematics. The English sentences were manually translated into French. All symbolic expressions were ''tokenized'', i.e. replaced with token words such as ''MATH66X''. All remaining LaTeX formatting is removed.

These three sub-corpora were combined into one heterogeneous corpus containing a total of 267,842 English-French parallel text chunks, which we refer to as the ''main'' corpus in the remainder of this paper. Most of the text chunks are sentences, so we refer to them as ''sentences'' hereafter. Each sentence pair in the main corpus was word-tokenized to ensure all tokens in a sentence including punctuation were separated by a space and treated individually during training. We randomly shuffled the main corpus, and then randomly split the text pairs into training (80%), validation (10%), and test (10%) sets.

For additional testing, we used a further custom corpus, the ''linear code'' corpus, containing 160 sentence pairs, extracted from mathematical course notes prepared by M. Nevins on linear error-correcting codes, a subject that did not appear in our main corpus. Beginning with French sentences extracted from the source document we manually translated the sentences to English, and removed LaTeX commands except within symbolic expressions. For all testing *except* the ''Google Raw'' variant (see details below), we tokenized all symbolic expressions.

B. PARSING METHODS

We use a modular design that decouples the LaTeX parsing and machine translation aspects. The first task is to parse the LaTeX document and extract all natural language text for translation while preserving enough document structure and LaTeX commands to reconstruct a full document. The second task, addressed in the next section, is to translate the extracted natural language text to French.

For the first task of parsing the LaTeX document, our main tool was the Pandoc Universal Document Converter [16], using the Python wrapper `pypandoc`. For our purposes, since we aim to translate an English LaTeX document to French while preserving the original LaTeX as much as possible, we are ''converting'' a document from LaTeX to LaTeX. Our purpose in doing so is to leverage an intermediate document representation internal to Pandoc, the JSON-formatted *abstract syntax tree (AST)*, and a mechanism for performing operations on this syntax tree: pandoc filters, implemented using the Python package `pandocfilters`. The abstract syntax tree is organized by block elements, such as paragraphs, bulleted lists, and tables, each of which contains a list of ''inline elements'' including strings of individual words, spaces, and math. Since the filters integrate into the Pandoc LaTeX-to-LaTeX file conversion, the entire translation process executes with one Pandoc function call. An overview of the process is shown here:

We use two pandoc filters, each of which modifies the abstract syntax tree. The first ''core'' Pandoc filter translates all block elements that contain natural language text, by joining strings of text and math symbols into whole sentences (or sometimes phrases, as in titles), translating those sentences (see next section), and putting the translated sentences back into the abstract syntax tree. Note that the larger-scaled block-based structure of the document is preserved in the abstract syntax tree.

Typically, pandoc filters act on individual inline elements such as strings. However, for translation of text, this is very limited, as each string containing an individual token would have to be translated separately, instead of a whole sentence. Thus, we combined these inline elements into whole sentences. We accomplished this in a ''string-joining'' function consisting of two layers: manipulating individual block elements through the pandoc interface, and further manipulating the inline elements directly. Importantly, we include mathematical formulas (inline or displayed) in our sentences, tokenized into ''MATH'' tokens of the form ''MATHnX'', where n is the index of the token for later retrieval of its corresponding mathematical formula. The original formulas are saved in a JSON object with their corresponding token name as a key. This way, full sentences containing these formulas can be translated without the loss of any important surrounding context.

For example, consider a simple mathematical sentence in LaTeX: ''Let Y have mean μ and variance σ^2, and an unknown p.d.f. p_Y that is everywhere nonzero.'' Within a JSON-structured paragraph block, this sentence would be

represented as a list of inline elements, the beginning of which is shown in Fig. 1. Without manipulating the mathematical formulas within this sentence, the sentence would be split up by these objects, so the largest possible concatenated phrases for translation would be "Let", "have mean", "and variance", etc. which would pose a major limitation for translation quality (as we demonstrate in Table 2 below). After tokenizing "math" objects, the whole sentence is concatenated into a single "Str" element: ["t":"Str","c":"Let MATH1X have mean MATH2X and variance MATH3X, and an unknown p.d.f. MATH4X that is everywhere nonzero."] The format of the math tokens is such that they are treated as unknown English words by the translation module(s), and left unchanged in the translated sentences. Mathematical expressions can now be treated as individual tokens during translation, and the entire sentence can be translated, optimizing translation quality. Note that after the filter runs, pandoc automatically converts the "Str" sentence object back into individual "Str" word and "Space" objects, as in the original abstract syntax tree.

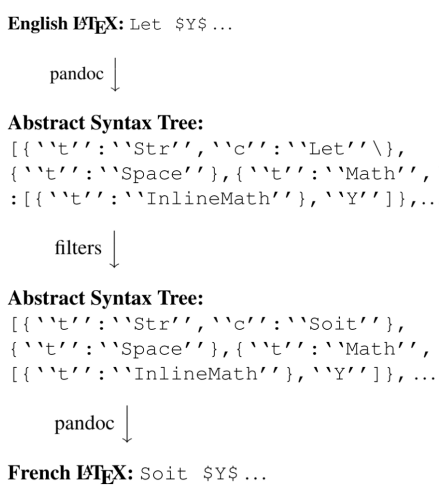

FIGURE 1. Data flow of LaTeX parsing module using the pandoc universal document converter. The internal representation is the *abstract syntax tree* (AST) which is in JSON format.

Once all natural language text in the modified LaTeX document has been translated to French, it remains to replace the "MATH" tokens with the original mathematical formulas, using the saved JSON object created by the "core" pandoc filter. This is accomplished by a second pandoc filter called the "detokenizer".

Finally, pandoc creates a new LaTeX file using the translated text and the structure of the original document. The entire process is represented in Fig. 2. Note that some changes to the LaTeX commands are introduced as a result of pandoc's abstract syntax tree not being able to completely represent all LaTeX commands. However when the original and translated documents are compiled to PDF format, very few differences in format are seen.

C. TRANSLATION METHODS

For translating text, we trained a custom neural machine translation (NMT) model using the Transformer architecture for neural sequence transduction introduced by Vaswani *et al.* [27]. The Transformer is a neural sequence transduction model, i.e. a "sequence-to-sequence" translator implemented using a neural network that outputs, for any position t in the output sequence y, a conditional distribution $p\left(y_t \middle| y_{<t}, x\right)$ based on the entire input sequence x and the preceding outputs $y_{<t}$. Like most other such models, the Transformer has an encoder-decoder structure. *Attention* mechanisms apply weights to elements of the input sequence, which vary according to the position t in the output sequence. This allows the network to "pay attention" to certain inputs, for example when producing the first word of an output sentence, the network may pay most attention to the first word of the input sentence. The key distinguishing feature of the Transformer model is its use of Multi-Head Self-Attention which allows it view the input sequence from different "points of view" by applying several parallel attention functions. For example, when encoding the word "kicked" in the sentence "I kicked the ball", one may pay more attention to the "I" or the "ball", corresponding to asking the questions "who performed the action?" or "what was kicked?" In this toy example, the first point of view would aid most in conjugating the corresponding output verb, while the second point of view would aid most in translating the verb root. A key feature of the Transformer attention mechanism is that it is highly parallelizable, allowing much faster training than previous comparable models. We trained a Transformer "base model" as implemented in OpenNMT [12], using the training subset of the main corpus described above. Training details are given in the next section.

The output of our trained Transformer model is already high-quality, as will be seen in the next section. However, while we can expect our model to have good performance on mathematical text (thanks to the inclusion of mathematical text in our training set), we do not expect it to perform as well on general English text as commercial translation services such as Google Translate, due mainly to our limited training set. For this reason, we used Google Translate, with the custom math glossary described above, as a "backup" translator in our final system, as follows.

We first run all sentences through our main Transformer model. The output of this process is not just a translated sentence but also a cumulative log conditional likelihood score $\sum_t \log p\left(y_t \middle| y_{<t}, x\right)$, where the sum is over all tokens in the output sentence. Dividing this by the length of the sentence gives the mean log conditional likelihood per token, which is a measure of how confident the system is of its prediction. This is commonly converted into a *perplexity* value, calculated as exp (−mean conditional likelihood), which is lowest for the most confident predictions. The median perplexity value for the Transformer model on the validation set is 1.68. We established a threshold value for perplexity, and whenever our Transformer model produced a sentence with perplexity

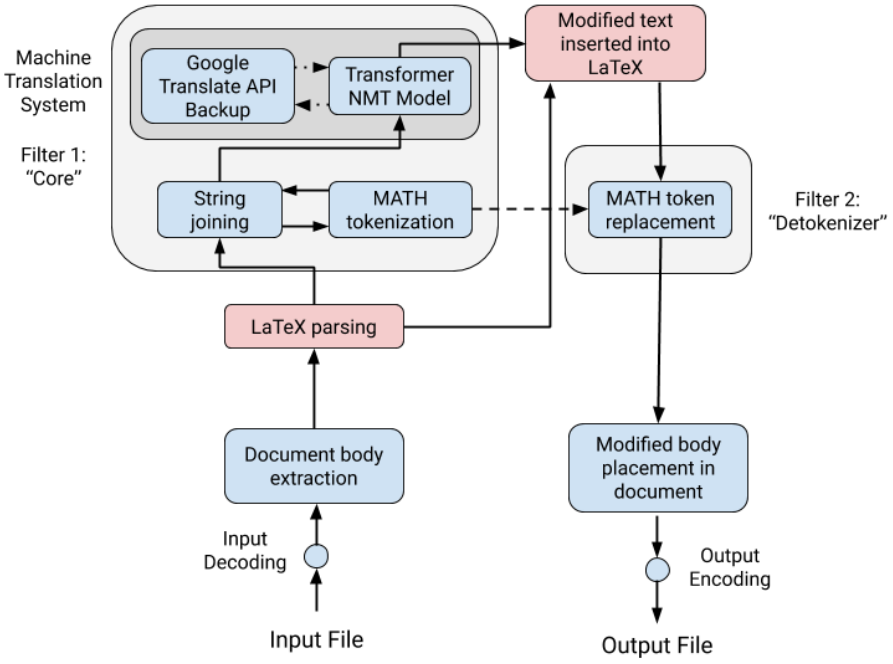

FIGURE 2. Overview of PolyMath Translator. The pink-coloured modules are implemented using the pandoc universal document converter, see also Fig. 1.

above this threshold, we discarded the output and instead re-translated the input sentence using the Google Translate API with our custom math glossary. We tuned the perplexity threshold to maximize the BLEU score (see next section) on the validation set, see Figure 3. The optimal threshold, which we used for all testing, was 1.75. Our use of a perplexity threshold to select a translation model may be considered to be an elementary form of ensemble learning.

D. FRENCH-SPECIFIC LaTeX MODIFICATIONS

Finally, since French LaTeX has its own typographical conventions, PolyMath adds a "french" option to the document class and adds the following lines to the document header:

```
\usepackage[T1]{fontenc}
\usepackage{babel}
```

It also explicitly translates all LaTeX-style double quotes (i.e. two consecutive backquotes or two consecutive single quotes) into `\og` and `fg{}` respectively.

E. TRAINING, VALIDATION AND TESTING

As noted earlier, our main corpus was randomly split into training (80%), validation (10%), and test (10%) sets. The training subset contains 214,272 English-French sentence pairs, while the validation and test subsets each contain 26,785 sentence pairs.

To evaluate the quality of machine translations as compared to given reference (i.e. target) translations, we used the BLEU metric [19] ("Bilingual Evaluation Understudy"), specifically the implementation in the sacreBLEU python package [21]. This metric is standard in the field and has been

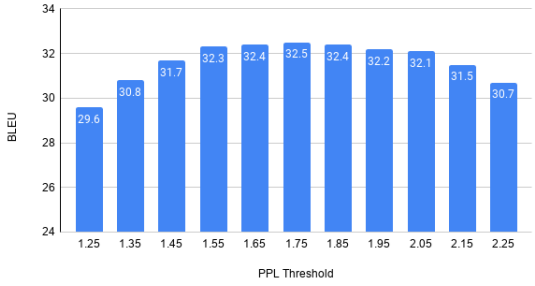

FIGURE 3. Results of tuning the perplexity ("PPL") threshold during validation. If the Transformer model produces a perplexity value above this threshold, then the input sentence is re-translated using Google Translate with our custom math glossary. The maximum BLEU is obtained for a threshold of 1.75.

shown to correlate well with human judgement of translation quality. We follow the most common convention and scale the metric to a score between 0 and 100, where 0 is the poorest quality (no overlap with the reference translations) and 100 is the highest quality.

The BLEU score takes into account the length of the translation output in comparison to the reference, as well as its "precision", through counting the number of matching unigrams, bigrams, trigrams, and four-grams within the output and reference translations. Specifically, the BLEU score is the product of a brevity penalty and an n-gram overlap score, each defined as follows:

$$\text{brevity penalty} = \min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right),$$

$$\text{n-gram overlap} = \left(\prod_{n=1}^{4} \text{precision}_n\right)^{\frac{1}{4}},$$

where each term "precision$_n$" is basically the proportion of the n-grams in the output that appear in the target. See [19] or [9] for details.

We trained the Transformer model with a vocabulary of 50,000 words on the training subset of the main corpus. We set the batch size to 3072 tokens and a maximum of 100,000 steps or 65 epochs of the entire corpus for training, and we used default values of all other hyperparameters. We did not tune any model hyperparameters. We used an early stopping condition, to avoid over-fitting and save computational cost once the model converges, for which we used the validation subset of the main corpus. The stopping condition was: stop when the validation BLEU score does not improve by more than 0.2 points over the last 4 evaluations. After 50,000 steps or 32 epochs, our model met the early stopping criteria and stopped training with a validation BLEU score of 28.7. This training required 12 hours on 8 Tesla V100 GPUs.

During testing, each sentence is first translated by the Transformer model, and if it has a test perplexity of less than or equal to a threshold value of 1.75, this prediction is retained, otherwise, the sentence is instead translated by Google Translate with our custom math glossary. The threshold value was tuned on the validation set as described in Section II-C. We note that during testing on the "linear code" corpus, the full PolyMath system used the Transformer NMT model for 71% of the sentences, and used the "backup system" of Google Translate with custom glossary for the other 29% of the sentences.

After training the Transformer model and integrating the other components described above, we evaluate the quality of the PolyMath system by computing mean BLEU scores for our two test corpora: the testing subset of the main multi-domain corpus; and the "linear code" corpus of mathematical text. We also perform ablation testing, i.e. we compare the entire PolyMath Translation system with versions in which certain of its features are disabled. Specifically, we examine two variants: "PM-Transformer", which includes the usual LaTeX parsing and math tokenization but then translates sentences using only our Transformer NMT model; and "PM-Google" which is similar but translates sentences using only Google Translate, with no custom glossary.

We performed further ablation testing on the linear code corpus. The "PM-Piece" model parses the LaTeX as usual except that the symbolic expressions are not tokenized; instead, translation is performed "piecewise", i.e. each sentence fragment between symbolic expressions (if any) is translated separately and then text fragments and symbolic expressions reassembled to make a full sentence. The BLEU scores are calculated on the full translated sentence. Finally, the "Google Raw" model is that each sentence is passed to Google Translate in its original form, which may include LaTeX commands inside of symbolic expressions but does not include any LaTeX commands for text or document formatting.

III. RESULTS

We first performed an exploratory comparison of our three sub-corpora of the OPUS Wikipedia corpus, corresponding to the domains: mathematics, geography, and sociology. As noted in Section II-A, these sub-corpora are not truly single-domain, but instead are multi-domain corpora in which the selected domain is heavily weighted. For each corpus, we compute: mean sentence length (in English), vocabulary (in English), and BLEU score for English-to-French translation using Google Translate (without a custom glossary). The sentence length is the total number of tokens (including punctuation). To avoid the confounding effects of corpus size and sentence size on vocabulary, the vocabulary reported is the number of unique tokens in first 100,000 tokens. The results are shown in Table 1. Sentences in the math corpus are shorter (mean 33.0 tokens) than in the other two corpora (mean 51.2 and 52.4), and the vocabulary is smaller (13.2K vs. 16.2K and 17.6K).

TABLE 1. Characteristics of three sub-corpora of the OPUS Wikipedia corpus. Each corpus is multi-domain, however the target domain is over-represented. Vocabulary is the number of tokens in the first 100,000 tokens in the corpus. The BLEU score is for English-to-French translation using Google Translate.

	sentence length	vocabulary	BLEU
Math	33.0	13,222	27.2
Geography	51.2	16,271	23.8
Sociology	52.4	17,591	24.5

Table 2 shows our main results, which are mean BLEU scores calculated on our two test corpora: the test subset of the multi-domain "main corpus"; and the "linear code" corpus of exclusively mathematical text. Note that both test corpora consist of sentence pairs already preprocessed, with any LaTeX commands outside of symbolic expressions removed. The results of testing on full LaTeX documents are reported later.

The highlight of these BLEU results is the score of 53.6 on mathematical text (the "linear code" corpus). This is much higher than the state-of-the-art for general multi-domain English-to-French machine translation, which is 41.8 [27].

The rest of Table 2 shows the results of ablation testing. On the "linear code" test corpus, using only the Transformer NMT model resulted in a 3.2 point drop in BLEU score; while using only Google Translate (with no custom glossary) resulting in a 7.1 point drop in BLEU score. It is also noteworthy that the Transformer model outperformed Google Translate, even on multi-domain text. This may be partly due to the similarity of training and test data, which gave the Transformer model an advantage. However it does suggest that our main corpus was large enough to support a comprehensive language model.

The further ablation testing showed large drops in performance when math tokenization was not used. The "PM-Piece" system, which broke sentences into fragments at each symbolic expression and translated the pieces separately, resulted in a BLEU score of 38.0 (a drop of 15.6 points). The "Google Raw" system, in which each sentence is

TABLE 2. Mean BLEU scores on two held-out test corpora: (first column) the "test" subset of our multi-domain main corpus; and (second column) the "linear code" corpus consisting entirely of mathematical text. The rows correspond to: (i) (Full PM) the full PolyMath Translator system; (ii) (PM-Transformer) the PolyMath system except only using the Transformer model; (iii) (PM-Google) the PolyMath system except only using Google Translate; (iv) (PM-Piece) the PolyMath system without math tokenization, translating only "piecewise", i.e. individually translating each sentence fragment between symbolic expressions (if any), but still calculating the BLEU score on whole sentences; (v) (Google Raw) Google Translate applied to "raw" sentences including some LaTeX commands in symbolic expressions but not including any LaTeX formatting commands.

	main corpus (multi-domain)	linear code corpus (mathematical)
Full PM	**32.5**	**53.6**
PM-Transformer	29.0 *(-3.5)*	50.4 *(-3.2)*
PM-Google	27.7 *(-4.8)*	46.5 *(-7.1)*
PM-Piece	—	38.0 *(-15.6)*
Google Raw	—	31.6 *(-22.0)*

submitted directly to Google Translate in its original form (including symbolic expressions but without LaTeX formatting commands) resulted in a BLEU score of 31.6, which is 22.0 points below our best method; a more informative comparison is that it is 14.9 points below "PM-Google", i.e. the PolyMath system using only Google Translate and not our Transformer model. Together these tests demonstrate that translation of full sentences with math tokenization results in a large improvement in translation quality.

In Table 3 we illustrate that the PolyMath Translator is a complete LaTeX document translation system, integrating natural language translation with LaTeX document parsing and French language support. A single sentence, containing LaTeX commands for symbolic expressions, text formatting and document formatting, is translated first by Google Translate and then by the PolyMath Translator. Unsurprisingly, the output of Google Translate is almost unusable: not only can it not be compiled since it not valid LaTeX, but since Google Translate is not LaTeX-aware, it mistakenly translates the LaTeX command "\in" into "\ dans". It also does not understand the LaTeX system for representing accented characters. In contrast, the PolyMath Translator directly gives the desired result: a piece of LaTeX code that compiles to give a perfect translation of the sentence with formatted text and symbolic expressions.

Finally, all of the LaTeX source documents used in this study, as well as the present manuscript in preprint form, were translated to French by the PolyMath Translator system, and the results compiled to PDF using TeXShop or Overleaf. Most of the output documents compiled to PDF without any manual editing required. The translated version of the present manuscript in preprint form is included in the Supplementary Material, compiled to PDF; this is the output of the PolyMath Translator, with minor manual edits made to only 6 lines of the output file (mostly related to figures). There are some slight differences in format between the original manuscript and the French version, due to the pandoc internal representation (the abstract syntax tree) not being a perfect representation of the LaTeX input. However, the formatting is very similar to the original paper, and the English to French translation is very good.

We have made the PolyMath Translator available as a web-service at `polymathtrans.ai`.

IV. DISCUSSION

We have implemented a prototype machine translation system for LaTeX documents containing mathematical text, the PolyMath Translator. The current implementation translates English LaTeX to French LaTeX, attaining a BLEU score of 53.6 on a held-out test corpus of mathematical sentences. (See Table 2.) This is much higher than the state-of-the-art for general multi-domain English to French machine translation, which is 41.8, as attained by the "big" Transformer model [27]. Further, it is comparable to other state-of-the-art BLEU scores on specialized domains; for example in [11], maximum BLEU scores for German-to-English translation domain-specific corpora were: 37.8 (law), 49.0 (medicine), and 59.4 (information technology).

Our ablation testing demonstrated that tokenization of symbolic expressions, i.e. the temporary conversion of an entire symbolic expression to a single token, was essential to obtaining high quality translations. Indeed, Table 2 shows that two alternative approaches: piecewise translation of sentences, and translation of "raw" entire sentences, lead to much lower mean BLEU scores.

This was not surprising to us, due to our earlier hypothesis that symbolic expressions usually (though certainly not always) function as nouns in mathematical text, in which case replacing them with a single token retains the grammatical structure of a sentence. In contrast, breaking a sentence into pieces at symbolic expressions and translating the pieces separately (as in the method "PM-Piece" in Table 2) seems very unpromising for sentences containing several symbolic expressions, since the resulting pieces are so small as to lose grammatical and semantic context.

Our main contribution is the implementation and evaluation of the PolyMath translation system, which includes LaTeX parsing, tokenization of symbolic expressions, and a Transformer-based model trained on a heterogeneous corpus containing a small proportion of domain-specific sentences. This system is available as a web-service at `polymathtrans.ai`.

Our secondary contributions are: an examination of how mathematical text differs from text in other domains; and the publication of the "linear code" corpus of English-French sentence pairs with symbolic expressions tokenized. [18]

Our exploratory comparison of text from different domains, summarized in Table 1, suggests that mathematical text has shorter sentences and a smaller vocabulary than text from other specialized domains. The finding of a small vocabulary is consistent with our anecdotal evidence mentioned in the Introduction, and also with our observation, noted in Section II-A that the vocabulary of our "math-wiki" subset of the OPUS Wikipedia corpus was approximately half of the vocabulary of other random multi-domain subsets of the same corpus of the same size. However the finding of shorter sentences is opposite to the tentative inference we drew in

TABLE 3. Illustration of two translations of English LaTeX into French. The top row shows an excerpt from a compiled LaTeX document. The corresponding LaTeX source code was entered into to two English-to-French translation systems: (i) Google Translate correctly translates the words, however is unaware of LaTeX syntax, resulting in an un-compilable document (e.g. "\in" is translated to "\ dans"); while (ii) PolyMath correctly translates the entire LaTeX document into a French LaTeX document that compiles without further editing.

Original English LaTeX :	**Definition 2.1.** Let $x \in \mathbb{F}^n$. The *closed ball of radius* r *centered at* x is $$S_r(x) = \{y \in \mathbb{F}^n \mid d(x, y) \leq r\}.$$

Google Translate:	`\ begin {defn}`
	`Soit $ x \ dans \ F ^ n$.}`
	`La \ define {boule fermÃl'e de rayon $ r $`
	`centrÃl'e sur $ x $} est`
	`$$`
	`S_r (x) = \ {y \ in \ F ^ n \ mid d (x, y) \ leq r \}.`
	`$$`
	`\ end {defn}`

PolyMath output LaTeX:	**Definition 2.1.** Soit $x \in \mathbb{F}^n$. La *boule fermée du rayon de* r *centrée à* x est $$S_r(x) = \{y \in \mathbb{F}^n \mid d(x, y) \leq r\}.$$

the Introduction from Yasseri *et al.*'s study of Wikipedia entries [32], which found that the most complex examples studied were from Philosophy and Physics. (Mathematics was not considered.) We also found (see Table 1) that Google Translate produced higher-quality translations of mathematical text than text from other domains, which is consistent with our findings of a smaller vocabulary and shorter sentence lengths in mathematical text, and supports the general conclusion that, from a natural language point of view, mathematical text is *simpler* than text in most other domains. The conclusion is broadly consistent with the finding of Jin *et al.* [11] (mentioned above) that BLEU scores for domain-specific corpora were higher for information technology than for law and medicine (if we consider that "information technology" overlaps with computer science and hence with mathematics).

Our own experiments with our PolyMath translation system also support the general conclusion that mathematical text is simpler and easier to translate than text in other domains. Indeed, as noted earlier, our BLEU score of 53.6 on mathematical text is higher than the state-of-the-art for general English to French machine translation, which is 41.8, attained by the "big" Transformer model [27]. Our lower BLEU score of 32.5 on our multi-domain main corpus (29.0 using the Transformer model only), is more in line with expectation, especially since we used the "base model" Transformer and only trained for 12 hours on 8 V100s, while the state-of-the-art result of 41.8 required training a "big" Transformer model for 3 days on 8 V100s. Comparing our BLEU results for the two testing corpora, we conclude that, since we used the same translation system, our increased score on mathematical text is due to the relative simplicity of this domain.

In summary, mathematical text seems to be simpler and easier to translate than text from most other domains. This offers an intriguing possibility of better-than-usual machine translation, and other natural language processing, in this domain, as suggested by our high BLEU score of 53.6 on a test corpus, using a prototype translation system that has evident room for improvement.

The PolyMath Translator produces LaTeX documents that can be compiled to PDF without further editing. Since French LaTeX has its own typographical conventions, PolyMath makes small changes to the header and quotation marks, as detailed in the Methods section. Once we added this step, we found that in all of our experiments the output of PolyMath compiled to PDF (using TeXShop) without error; thus PolyMath is robust in its handling of LaTeX syntax.

Several obvious opportunities exist for improving this system. We can expand and improve our corpora and glossary, using both automated tools, for example automatic multi-word terminology extraction [28], [30], and manual proofreading. This might involve the use of large document collections including `arXiv`, which includes some publicly available LaTeX sources, see [5]. We can incorporate recent advances in machine translation research in areas including deep learning [20], [34], Bayesian modelling [33], and methods for incorporating whole document context [15]. Further, given the relative ease of machine translation in the mathematical text domain, and at the same time, the lack of curated corpora of mathematical sentence pairs, we are optimistic that semi-supervised translation [3], [11] and multi-language models [7] will be successful in this domain.

Also, we will investigate alternative methods of LaTeX parsing that will allow us to exactly retain all of the LaTeX commands in the original document, rather than passing them through the pandoc internal representation, which introduces some changes as mentioned above.

An intriguing possibility specific to the domain of mathematical text is to improve translations by semantic understanding of the content of mathematical formulas, see e.g. [10], [13]. A simpler early target would be to classify the

formulas by their parts of speech, usually nouns, pronouns or clauses; and encode them in a way that is usable by the translation modules.

While the future research possibilities are exciting, one of our main conclusions is that high-quality translation of text in mathematics and neighbouring domains, in LaTeX format, is possible *now*, without waiting for future research breakthroughs. We hope that automatic translation of LaTeX articles between multiple languages soon becomes standard practice.

ACKNOWLEDGMENT

The authors thank Blair Drummond for suggesting the use of pandoc, and Monica Nevins for contributing the source material for our ''linear code'' corpus. Thanks also to Othmane Ayoub for related discussions, and to the anonymous reviewers for their comments.

REFERENCES

[1] American Sociological Association. *Introduction to Sociology: Glossary*. Accessed: Jan. 30, 2021. [Online]. Available: https://www.asanet.org/sites/default/files/savvy/introtosociology/Documents/Glossary.html
[2] D. Britz, Q. Le, and R. Pryzant, ''Effective domain mixing for neural machine translation,'' in *Proc. 2nd Conf. Mach. Transl.*, 2017, pp. 118–126.
[3] Y. Cheng, ''Semi-supervised learning for neural machine translation,'' in *Joint Training for Neural Machine Translation*. Singapore: Springer, 2019, pp. 25–40.
[4] C. Chu and R. Wang, ''A survey of domain adaptation for neural machine translation,'' in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1304–1319.
[5] C. B. Clement, M. Bierbaum, K. P. O'Keeffe, and A. A. Alemi, ''On the use of ArXiv as a dataset,'' 2019, *arXiv:1905.00075*. [Online]. Available: http://arxiv.org/abs/1905.00075
[6] Alberta Education. *Lexique de Mathématiques*. Accessed: Sep. 20, 2020. [Online]. Available: https://education.alberta.ca/media/481794/en-lexique-math-juillet2015.pdf
[7] O. Firat, K. Cho, and Y. Bengio, ''Multi-way, multilingual neural machine translation with a shared attention mechanism,'' 2016, *arXiv:1601.01073*. [Online]. Available: http://arxiv.org/abs/1601.01073
[8] Ulrich Germann. *Aligned Hansards of the 36th Parliament of Canada, Release 2001-1a*. Accessed: Sep. 20, 2020. [Online]. Available: https://www.isi.edu/natural-language/download/hansard
[9] Google. *Evaluating Models*. Accessed: Sep. 25, 2020. [Online]. Available: https://cloud.google.com/translate/automl/docs/evaluate
[10] A. Greiner-Petter, A. Youssef, T. Ruas, B. R. Miller, M. Schubotz, A. Aizawa, and B. Gipp, ''Math-word embedding in math search and semantic extraction,'' *Scientometrics*, vol. 125, no. 3, pp. 3017–3046, Dec. 2020.
[11] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, ''A simple baseline to semi-supervised domain adaptation for machine translation,'' 2020, *arXiv:2001.08140*. [Online]. Available: http://arxiv.org/abs/2001.08140
[12] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, ''OpenNMT: Open-source toolkit for neural machine translation,'' in *Proc. ACL, Syst. Demonstrations*. Vancouver, BC, Canada: Association Computational Linguistics, 2017, pp. 67–72.
[13] G. Y. Kristianto, G. Topic, and A. Aizawa, ''Extracting textual descriptions of mathematical expressions in scientific papers,'' *D-Lib Mag.*, vol. 20, nos. 11–12, p. 9, Nov. 2014.
[14] S.-C. Loh, L. Kong, and H.-S. Hung, ''Machine translation of Chinese mathematical articles,'' *ALLC Bull.*, vol. 6, no. 2, pp. 111–120, 1978.
[15] V. Macé and C. Servan, ''Using whole document context in neural machine translation,'' 2019, *arXiv:1910.07481*. [Online]. Available: http://arxiv.org/abs/1910.07481
[16] J. MacFarlane. *Pandoc Universal Document Converter*. Accessed: May 30, 2020. [Online]. Available: https://pandoc.org/
[17] M. Nagao, J.-I. Tsujii, K. Yada, and T. Kakimoto, ''An English Japanese machine translation system of the titles of scientific and engineering papers,'' in *Proc. 9th Conf. Comput. Linguistics*, 1982, pp. 1–8.
[18] T. Schmah and A. Ohri, ''Linear code sentences English/French,'' *IEEE Dataport*, Oct. 2020, doi: 10.21227/9h6z-z514.
[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ''BLEU: A method for automatic evaluation of machine translation,'' in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.
[20] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtský, ''Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals,'' *Nature Commun.*, vol. 11, no. 1, pp. 1–15, Dec. 2020.
[21] M. Post, ''A call for clarity in reporting BLEU scores,'' in *Proc. 3rd Conf. Mach. Transl., Res. Papers*, 2018, pp. 186–191.
[22] E. Reyssat. *Traduction de Quelques Termes Mathématiques*. Accessed: Aug. 30, 2020. [Online]. Available: https://perso.crans.org/besson/publis/dicoAF.html
[23] F. Schweiger, ''The grammar of mathematical symbolism,'' in *Proc. Evelyne Barbin, Nad'a StehlíKová und Constantinos Tzanakis eds History Epistemol. Math. Educ. Proc. 5th European Summer Univ.*, 2008, pp. 423–430.
[24] American Mathematical Society. *Mathscinet Mathematical Reviews*. Accessed: Oct. 1, 2020. [Online]. Available: https://mathscinet.ams.org/
[25] A. R. Terryn, V. Hoste, and E. Lefever, ''In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora,'' *Lang. Resour. Eval.*, vol. 54, pp. 385–418, Mar. 2019.
[26] J. Tiedemann, ''Parallel data, tools and interfaces in OPUS,'' in *Proc. Lrec*, 2012, pp. 2214–2218.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[28] B. Šandrih, C. Krstev, and R. Stankovic, ''Two approaches to compilation of bilingual multi-word terminology lists from lexical resources,'' *Natural Lang. Eng.*, vol. 26, no. 4, pp. 455–479, Jul. 2020.
[29] Q. Wang, C. Brown, C. Kaliszyk, and J. Urban, ''Exploration of neural machine translation in autoformalization of mathematics in Mizar,'' in *Proc. 9th ACM SIGPLAN Int. Conf. Certified Programs Proofs*, Jan. 2020, pp. 85–98.
[30] K. Wołk and K. Marasek, ''Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs,'' *Procedia Technol.*, vol. 18, pp. 126–132, Sep. 2014.
[31] M. Wolska, M. Grigore, and M. Kohlhase, ''Using discourse context to interpret object-denoting mathematical expressions,'' in *Proc. CICM Workshop DML*. Brno, Czech Republic: Masaryk Univ. Press, 2011, pp. 85–101.
[32] T. Yasseri, A. Kornai, and J. Kertész, ''A practical approach to language complexity: A Wikipedia case study,'' *PLoS ONE*, vol. 7, no. 11, Nov. 2012, Art. no. e48386.
[33] L. Yu, L. Sartran, W. Stokowiec, W. Ling, L. Kong, P. Blunsom, and C. Dyer, ''Better document-level machine translation with Bayes' rule,'' *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 346–360, Jun. 2020.
[34] B. Zhang, D. Xiong, J. Xie, and J. Su, ''Neural machine translation with GRU-gated attention model,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4688–4698, Nov. 2020.

ADITYA OHRI is a high-school senior in Toronto, Canada. And also a researcher in applied machine learning with interests in computational linguistics, machine translation, and computer vision. He is particularly interested in developing practical applications that aim to solve real-world problems, using his experience in software engineering. As an active advocate for STEM education within the community, he engages youth in data science through sports with a unique platform called Court Science.

TANYA SCHMAH received the Ph.D. degree from École Polytechnique Fédérale de Lausanne, Switzerland, in 2001. After positions at Mathematics departments with the University of Warwick, U.K., and Macquarie University, Australia, she was a Postdoctoral Fellow and Research Associate with the Department of Computer Science, University of Toronto, Canada, and a Program Manager of the Neuroinformatics Research Group, Rotman Research Institute Baycrest, Canada. In 2015, she joined the Faculty of the University of Ottawa, Canada, where she is currently an Associate Professor with the Department of Mathematics and Statistics. Her research interests include machine learning, neuroimage analysis, and geometric mechanics.