BMC Medical Informatics and Decision Making

DATABASE

Open Access

ParaMed: a parallel corpus for English–Chinese translation in the biomedical domain

Boxiang Liu[1*] and Liang Huang[1,2]

Abstract

Background: Biomedical language translation requires multi-lingual fluency as well as relevant domain knowledge. Such requirements make it challenging to train qualified translators and costly to generate high-quality translations. Machine translation represents an effective alternative, but accurate machine translation requires large amounts of in-domain data. While such datasets are abundant in general domains, they are less accessible in the biomedical domain. Chinese and English are two of the most widely spoken languages, yet to our knowledge, a parallel corpus does not exist for this language pair in the biomedical domain.

Description: We developed an effective pipeline to acquire and process an English-Chinese parallel corpus from the New England Journal of Medicine (NEJM). This corpus consists of about 100,000 sentence pairs and 3,000,000 tokens on each side. We showed that training on out-of-domain data and fine-tuning with as few as 4000 NEJM sentence pairs improve translation quality by 25.3 (13.4) BLEU for en→zh (zh→en) directions. Translation quality continues to improve at a slower pace on larger in-domain data subsets, with a total increase of 33.0 (24.3) BLEU for en→zh (zh→en) directions on the full dataset.

Conclusions: The code and data are available at https://github.com/boxiangliu/ParaMed.

Keywords: Machine translation, Natural language processing, Text mining

Background

Biomedical translation is used across various life science disciplines. Example applications include translation of clinical trial consent forms, regulatory documents, and interpretation within point-of-care facilities [1, 2]. Biomedical translation requires up-to-date domain knowledge and fluency in the source and target languages. Such requirements make it challenging to train qualified translators and costly to generate high-quality translations.

Recent advances in machine translation have demonstrated translation quality arguably on par with professional human translators in select domains [3]. Supervised training of machine translation models usually benefits from large amounts of parallel corpora and such effect is the most evident for neural machine translation models. However, the collection and alignment of parallel corpora requires significant time and labor, and such datasets are not available for all domains or language pairs.

Machine translation in the biomedical domain is characterized by a long tail of medical terminology. For example, the Unified Medical Language System (UMLS) developed by the National Institute of Health contains over 2 million names for over 900,000 concepts [4], much larger than the set of common English words. Therefore, domain adaptation (training on out-of-domain data and testing on in-domain data) from the general domain to the biomedical domain is challenging.

Two prevailing challenges impact biomedical translation quality when training is done on general-domain data. Biomedical concepts unseen in the general-domain training set (covariate shift) are difficult to translate

*Correspondence: jollier.liu@gmail.com
[1] Institute of Deep Learning, Baidu Research, Sunnyvale, USA
Full list of author information is available at the end of the article

accurately. Most medical terminologies, such as the word "oncogenesis", falls into this category. Additionally, concepts that appear in both biomedical domain and general domain but with different semantics present a second challenge. For example, "*primary* care" is translated to Chinese as "初级诊疗" whereas "*primary* element" is translated as "主要元素".

Various domain adaptation techniques have been developed. Synthetic data generation such as forward and backward translation [5] aims to augment out-of-domain parallel data with monolingual in-domain data. Data selection methods aim to select in-domain examples from general domain data [6]. Fine-tuning with a small amount of in-domain data has been shown to substantially improve translation quality [7].

While the need for biomedical parallel corpora is evident, they are not available for all language pairs. In a literature survey, we found that existing public biomedical parallel corpora are between European Languages (Table 1). The UFAL Medical Corpus covers language pairs from English to Czech, German, Spanish, French, Hungarian, Polish, Romanian and Swedish. The ReBEC dataset [8] contains Portuguese and English parallel texts obtained from 1,188 clinical trial documents in the Brazilian Clinical Trials Registry. The 2020 Conference on Machine Translation (WMT20) Biomedical Translation Workshop [9] provides training sentence pairs from Medline abstracts between English and Spanish/German/Portuguese/French/Italian/Russian, but only test sentence pairs for English and Chinese. The Khresmoi dataset [10] samples 1,500 English sentences from medical documents. These sentences are manually translated into Czech, French, German, Hungarian, Polish, Spanish, and Swedish. The MeSpEn dataset [11] contains English and Spanish parallel text collected from IBECS (Spanish Bibliographical Index in Health Sciences), SciELO (Scientific Electronic Library Online), Pubmed and MedlinePlus. Furthermore, we found that existing public English-Chinese parallel corpora are outside of the biomedical domain. The OPUS corpora contain English-Chinese translation from numerous sources such news, speeches, and movie subtitles [12]. Perhaps the most closely related is the UM-corpus. It contains parallel text from eight different domains, one of which is science and technology [13].

Table 1 Existing parallel corpus in the biomedical domain contains only European languages

Corpus	Language Components
UFAL	cs, de, en, es, fr, hu, pl, ro, sv
ReBEC	en, pt
WMT19	de, en, es, fr, pt
Khresmoi	cs, de, en, es, fr, hu, pl, sv
MeSpEn	en, es

cs: Czech, de: German, en: English, es: Spanish, fr: French, hu: Hungarian, pl: Polish, ro: Romanian, sv: Swedish

The New England Journal of Medicine (NEJM) provides Chinese translations of its publications dating back to 2011 (http://nejmqianyan.cn/). The website repository currently hosts nearly 2,000 articles, with new articles added weekly. These articles include original research articles, clinical case reports, review articles, commentaries, Journal Watch (viz. article highlights), etc. The articles are translated by professional translators and proofread by members of the NEJM editorial team. For research articles, translations on statistical analysis are proofread by statisticians who are native Chinese speakers.

In this study, we present an English–Chinese parallel corpus in the biomedical domain constructed from NEJM (Fig. 1). We provide sentence-aligned bitext for 1966 article pairs, totaling 97,441 sentence pairs. Further, we show that training a baseline model with the 2018 Conference on Machine Translation (WMT18) newswire data [14] and fine-tuning the model with the ParaMed dataset will significantly improve translation quality over the baseline model, suggesting that the ParaMed dataset will be useful in improving biomedical translation quality.

Our contributions are the following:

- We present the first English-Chinese parallel corpus in the biomedical domain. We only use the *open-access* portion of NEJM articles to comply with their editorial policy.
- We provide an end-to-end pipeline for constructing parallel corpus using web-crawled text. We compare several software packages for sentence boundary detection and alignment and provide guidelines on their performance in the biomedical domain.
- We show that fine-tuning on as few as 4,000 sentence pairs from ParaMed can improve translation quality by 25.3 (13.4) BLEU for en→zh (zh→en). Translation quality continues to improve at a slower pace on larger datasets, finishing at an increase of 33.0 (24.3) BLEU for en→zh (zh→en) on the full dataset.

Construction and content

Standard approaches to parallel corpus construction

Construction of a sentence-aligned parallel corpus from multilingual websites involves the following steps.

1 Documents in desired languages are crawled from multi-lingual websites.

English	Chinese
Tiotropium in Asthma Poorly Controlled with Standard Combination Therapy	噻托溴铵治疗标准联合疗法控制不佳的哮喘
Abstract	摘要
Background	背景
Some patients with asthma have frequent exacerbations and persistent airflow obstruction despite treatment with inhaled glucocorticoids and long-acting beta-agonists (LABAs).	一些哮喘患者尽管接受了吸入型糖皮质激素和长效β受体激动剂（LABA）治疗，仍有频繁的病情加重和持续气流阻塞。

Fig. 1 An overview of the ParaMed corpus construction. The input is the NEJM website and the output is a Chinese/English parallel corpus

2 Plain texts are extracted from crawled documents and normalized to remove special characters.
3 Documents from both languages are matched according to their contents.
4 Within each document, paragraphs are broken down into individual sentences.
5 Sentences are subsequently aligned into sentence pairs.
6 Aligned sentence pairs are filtered to remove duplicated and low-quality pairs.

While the first two steps are well-established engineering tasks, the last four are under active research. For step 3, the 2016 Conference on Machine Translation (WMT16) hosted a shared task for bilingual document alignment [15], in which the best entry relied on matching distinct bilingual phrase pairs [16]. For step 4, Read *et al.* [17] systematically evaluated nine existing tools for sentence boundary detection, among which `LingPipe` [18] and `Punkt` [19] are the top performers in the biomedical domain. Sentence alignment (step 5) is arguably the most challenging step. Compared with document alignment, sentence alignment uses a smaller amount of text but has more permutations. Various methods have been proposed, among which are length-based algorithm [20], lexicon-based algorithm [21–23], and translation-based algorithm [24, 25], with no consensus on the best performer. For step 6, WMT18 hosted a shared task on parallel corpora filtering [26], in which the best performer used dual conditional cross-entropy filtering [27].

The NEJM website provides hyperlinks between Chinese and English article pairs, allowing us to skip document alignment (step 3). Otherwise, we followed the best practices outlined therein and adapt them to our project (Fig. 2).

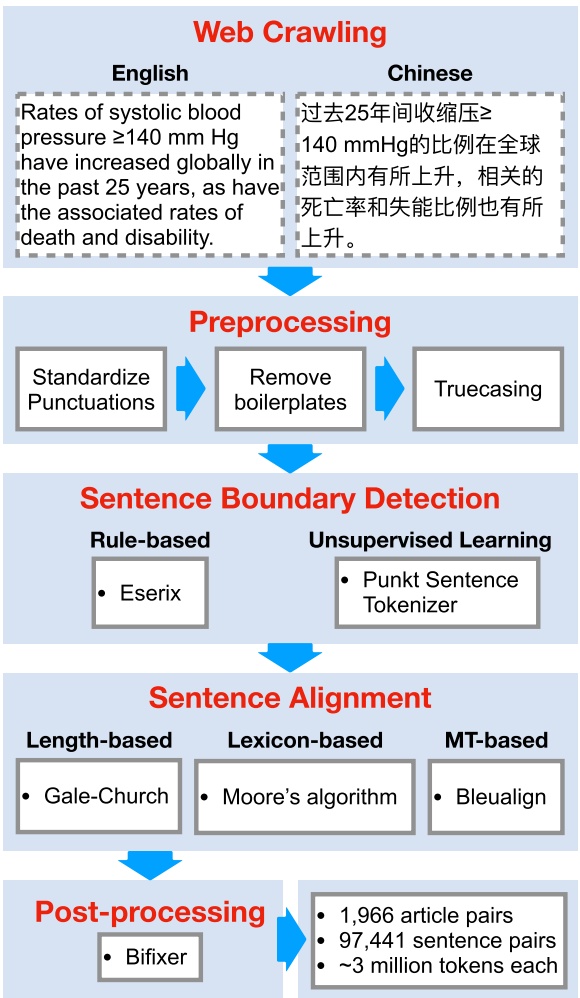

Fig. 2 The overall pipeline to construct the ParaMed dataset. NEJM webpages were crawled using Selenium. Various preprocessing steps were carried out to standardize punctuations and remove boilerplate texts. We tested two methods for splitting paragraphs into sentences, and three methods to align English and Chinese sentence into translated pairs. Duplicated sentence pairs were removed at the end

The New England Journal of Medicine Dataset

The Chinese website of the New England Journal of Medicine (https://www.nejmqianyan.cn/) provides open-access Chinese translations dating back to 2011. All articles were translated sentence for sentence by professional translators, with occasional sentence concatenation and division for fluency. In other words, one English sentence can be split into two or more Chinese sentences and vice versa. Translations were proofread by members of the editorial team and research articles were additionally proofread by statisticians. The Chinese translations are organized chronologically, making the content easy to crawl. Correspondent article pairs are connected via hyperlinks, eliminating the need for document alignment.

Web crawling

We used Selenium [28] to crawl all available Chinese and English articles. While paragraph orderings are maintained across languages, locations of display items—figures, tables, and associated captions—are shuffled. We removed display items to keep content orders identical across English and Chinese. The English NEJM website contains untranslated auxillary contents such as job boards and visual advertisements. We instructed Selenium to ignore auxillary contents as these interjections make sentence alignment challenging. Chinese NEJM translations are cleaner but contain boilerplate sentences such as names of translators.

These boilerplate contents were removed during preprocessing.

Preprocessing

We truecased letters and standardized punctuations for crawled articles with `moses` [29], and subsequently performed stitching and filtering described below.

Stitching incorrectly split sentences

A single sentence is occasionally split incorrectly due to inappropriate HTML tags. In Chinese articles, we found that sentence breaks can be inserted by mistake before citations and before punctuations. To stitch them, we assigned any text segment consisted only of citations and/or punctuations to its preceding sentence. For English, we noticed that the phrase "open in new tab" always incorrectly break a full sentence into two halves. We concatenated flanking sentences and remove the said phrase.

Filtering

Because display items and references are untranslated, we filtered out the following content for both languages:

- Figures and figure captions
- Tables and table legends
- Reference sections

Further, we removed content specific for either language. For Chinese, we removed any information about translators. For English, we removed:

- Video
- Interactive graphic
- Audio interview
- Visual abstract
- Quick take (video summary)

Sentence boundary detection (SBD)

Chinese sentences are concluded by three full-stop punctuations {！，？，。}. These punctuations are used exclusively for sentence separation. Unlike European languages, they do not double as decimal points or other linguistic markers. Further, Chinese quotation marks appear before sentence breaks, making it easy to detect sentence boundaries. Breaking English sentences is more challenging due to punctuation overloading.

Read et al. [17] showed that `punkt`, an unsupervised sentence tokenizer, is a top performer on biomedical corpora. We trained `punkt` on our ParaMed corpus and used the learned parameters to break sentences. Since `punkt` does not support the Chinese language, we implemented a custom regex-based tokenizer to split Chinese paragraphs into sentences.

Further, we tested a rule-based system `eserix` [30] designed to process the United Nations parallel corpus and has built-in support for both Chinese and English [31]. However, the default rules do not include commonly used abbreviation in biomedical literature, such as the word "Vol." as an abbreviation for "Volume". We added rules into the `eserix` ruleset specifically for the ParaMed corpus.

Sentence alignment

While many methods have been proposed for sentence alignment, there is no consensus on their performance in the biomedical domain. We focused on three main classes of methods: length-based, lexicon-based, and translation-based methods. We drew one method from each class: Gale-Church (length-based), Microsoft Aligner (lexicon-based), and Bleualign (translation-based). The Gale-Church algorithm finds sentence pairs based on the assumption that the lengths of source and target sentences should be similar [20]. The Microsoft Aligner integrates word correspondence with sentence lengths to search for sentence pairs [21]. Bleualign compares original and translated texts to search for anchor sentences and subsequently aligns the rest with the Gale-Church algorithm [25]. To compare these methods, we established a test set by manually aligning 1,019 sentence pairs from 12 articles. Table 2 shows the distribution of alignment types. Nearly 95% of all alignments are one-to-one. An example of one-to-many alignment is shown in Table 3.

Post-processing

Medical literature is highly structured. Certain sections such as the abstract, introduction, methods, results and discussion are almost universal across articles. We removed duplicated header and other repeated text with bifixer [32].

Table 2 Alignment counts in manually aligned sentence pairs, in which the majority are 1–1 alignments

zh-en	Count	Percent
0–1	10	1.0%
1–0	11	1.1%
1–1	964	94.6%
1–2	17	1.7%
2–1	15	1.5%
2–2	1	0.1%
2–3	1	0.1%

Table 3 An example 1-to-2 alignment for clause breaking. The red text denotes the English clause corresponding to the first Chinese sentence. Sotagliflozin is cited once in the English sentence, but repeated in two Chinese sentences

(a) Chinese		*shì*	*yìzhǒng*	*kǒufú*	*nà*	*pútaotáng*	*xiétóngzhuǎnyùndànbái*	*yī*	*hé*	*èr*	*de*	*yìzhìjì*	*wǒmén*		
	Sotagliflozin	是	一种	口服	钠–	葡萄糖	协同转运蛋白–	1	和	2	的	抑制剂。	我们		
	Sotagliflozin	**is**	**an**	**oral**	**sodium-**	**glucose**	**cotransporters-**	**1**	**and**	**2**	**'s**	**inhibitor**	**we**		
	píngjiàle	*zài*	*yīxíng*	*tángniàobìng*	*huànzhě*	*zhōng*	*liányòng*	*yídǎosù*	*hé*		*de*	*ānquánxìng*	*hé*	*liáoxiào*	
	评价了	在	1型	糖尿病	患者	中	联用	胰岛素	和	sotagliflozin	的	安全性	和	疗效。	
	evaluated	**in...**	**type-1**	**diabetes**	**patients**	**...in**	**combination**	**insulin**	**and**	**sotagliflozin**	**'s**	**safety**	**and**	**efficacy.**	
(b) English	We evaluated the safety and efficacy of sotagliflozin, an oral inhibitor of sodium–glucose cotransporters 1 and 2, in combination with insulin treatment in patients with type 1 diabetes.														

Training, development and test split

We selected 2102 sentence pairs from 39 latest articles as the test set and 2036 sentence pairs from the next latest 40 articles as the development set. The remaining 93,303 sentence pairs constitute the training set. To avoid data leakage, all sentences from each articles must be in one of either train, development, and test set.

Model architecture

We used the transformer model [33] in OpenNMT with 6 layers, each with an output size of 512 hidden units [34]. We used 8 attention heads and sinusoidal positional embedding. The final hidden feed-forward layer is of size 2,048. In addition, we used an LSTM [35] in OpenNMT with 512 hidden units.

Hardware and training procedure

We trained baseline transformer and LSTM models on the English-Chinese parallel corpus from WMT18 [36] consisting about 24.8 million sentence pairs. Sentences are encoded with Byte-Pair Encoding [37] with vocabularies of 16,000 tokens for each language. Sentence lengths are capped at 999 tokens, enough to accommodate most sentences. We trained these models on 8 Nvidia TitanX GPUs. For the transformer model, we used the Adam optimizer [38] with $lr = 2$, $\beta_1 = 0.9$, $\beta_2 = 0.997$ and 10,000 warm-up steps. We applied dropout with $p_d = 0.1$ and label smoothing with $\epsilon_{ls} = 0.1$. The model was trained for 500,000 steps in total. The training procedure took 4.5 days. We fine-tuned the baseline model on ParaMed for 100,000 steps with identical parameters. To establish a second comparison, we trained a transformer model *de novo* on the ParaMed corpus. For the LSTM model, we used the Adam optimizer with $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and label smoothing with $\epsilon_{ls} = 0.1$.

Utility and discussion

Statistics on crawled articles

The earliest official translation by NEJM dates back to 2011, and the number of translated articles has been on the rise year over year. Journal Watch (article highlights) leads in the number of articles, followed by original research and review articles. Figure 3 shows the distribution of articles by year and type.

Comparing pre- and post-filtered corpora

To remove untranslated text and display items, we manually compared the corresponding Chinese and English articles, identified HTML divisions to be filtered, and implemented a rule-based system to automatically filtered out matching HTML divisions ("Filtering" section). Figure 4 compares the number of Chinese and English paragraphs in each article pair before and after filtering. Before filtering, the number of Chinese paragraphs exceeds that of English for numerous articles, indicated by the grey sub-diagonal cloud. This is due to the various untranslated and boilerplate texts within the articles. The number of English and Chinese paragraphs in each article become closer after filtering.

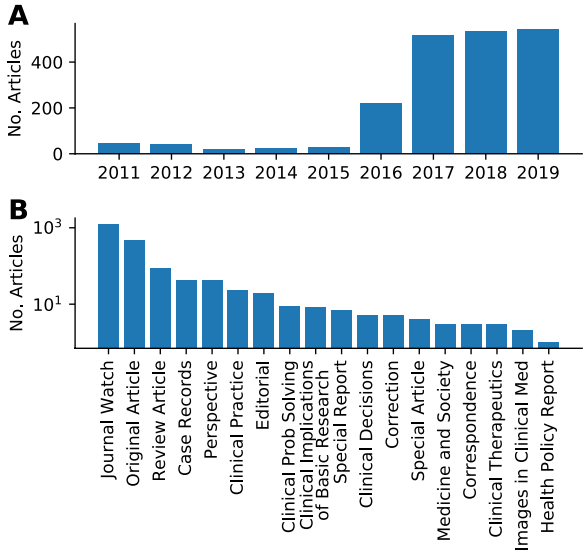

Fig. 3 Distribution of articles by year and by article type. **A** The number of translated articles has been on the rise since the year 2015. **B** Journal Watch, original articles, and review articles are the top three article types

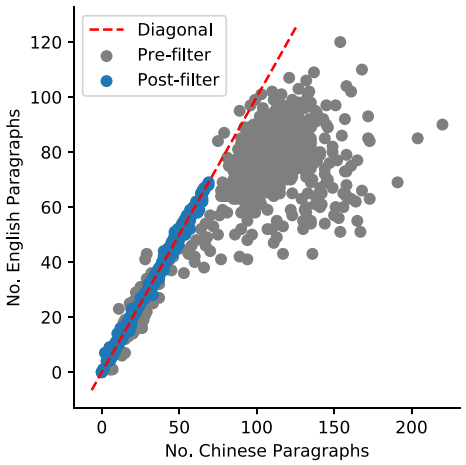

Fig. 4 Number of Chinese and English paragraphs are closer post-filtering compared to pre-filtering. Each point represents an English/Chinese article pair

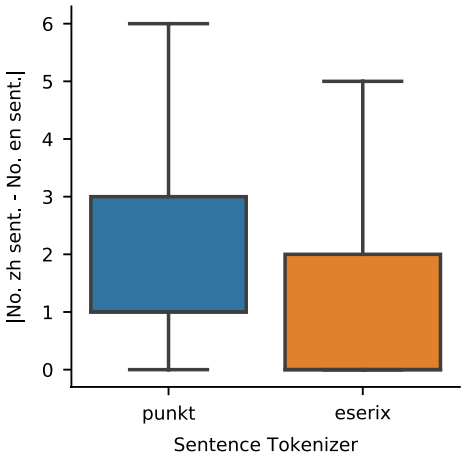

Fig. 5 The difference in the number of Chinese and English sentences are smaller in eserix than in punkt. Note that the distributions are left-skewed and the medians overlap with the 1st quartiles

Comparing sentence boundary detection algorithms

Because no systematic evaluation exists for sentence boundary detection in the biomedical domain, we tested two popular algorithms, `punkt` and `eserix`. To compare the two, we plotted the difference in the number of sentences. Because NEJM articles were translated sentence for sentence, the ideal SBD result should have a difference of zero. We found that difference is smaller for `eserix` (median difference = 0) than `punkt` (median difference = 1) and thus used it for downstream analysis (Fig. 5).

The two most frequent errors made by `punkt` were the failure to break at citations (Table 4) and erroneous breaks before open parentheses (Table 5). The latter created difficulty for sentence alignment because the Chinese sentence breaks appear after the close parenthesis. Conversely, `eserix` did not make these mistakes.

Comparing sentence alignment algorithms

To find correspondence between English and Chinese sentences, we tested three types of aligners, Gale-Church (length-based), Microsoft Aligner (lexicon-based), and Bleualign (translation-based), using a manually annotated set of 1,019 sentence pairs ("Sentence alignment" section). It should be noted that Bleualign was tested in both unidirectional (zh→en) and bidirectional (zh↔en) modes. The unidirectional mode has higher recall but lower precision, whereas the bidirectional mode has lower recall but higher precision. The majority of sentence pairs are one-to-one aligned (Table 2) and the performance of all algorithms degrade significantly for one-to-many and many-to-many alignments. Therefore, we focused on one-to-one alignments for this study. The precision, recall, and F1 scores are shown in Fig. 6. The Microsoft Aligner achieved the best F1 score and was used for downstream analysis.

Statistics of the ParaMed corpus

After the aligned sentences cleaned with bifixer [39], the final corpus contains 1,966 article pairs with a total of 97,441 sentences. We tokenized English sentences with moses [29] and Chinese sentences with Jieba. The English corpus contains 3,028,434 tokens and 55,673 unique tokens. The Chinese corpus contains 2,916,779 tokens and 46,700 unique tokens. All statistics are reported in Table 6.

Machine translation performance

To measure the effect that the ParaMed corpus has on medical translation, we compared the baseline transformer model trained on the WMT18 English-Chinese dataset and a fine-tuned model with the ParaMed corpus ("Hardware and training procedure" section). Although translations are evaluated bidirectionally,

Table 4 Example of a failure to break two English sentences due to a citation (red text). Corresponding Chinese sentences do not suffer from this problem

	shàngwèi	*bàodǎo*	*jīngyànzhèngde*	*yǒu*	*quánjīyīnzǔ*	*xiǎnzhùxìngde*	*jīyīnzuò*	*wèile*	*kèfú*	*yàngběnliàngde*	*xiànzhì*
(a) Chinese	尚未	报道	经验证的	有	全基因组	显著性的	基因座[12−14]。	为了	克服	样本量的	限制...
	have not	**been reported**	**replicated**	**with**	**genome-wide**	**significance**	**loci**	**To**	**overcome**	**sample-size**	**limitations...**
(b) English	No replicated loci with genome-wide significance have been reported.[12−14] To overcome sample-size limitations...										

Table 5 Example of an erroneous break before the blue text. Notice the additional period before the open parenthesis for the English text

	yǔ	*ānwèijì*	*zǔ*	*xiāngbǐ*	*fúxiè*	*zài*	*pàtuōzhūdānkàng*	*zǔ*	*jiàowéi*	*chángjiàn*	*yóu*	*huòfūmàn*	*luóshì*
(a) Chinese	与	安慰剂	组	相比，	腹泻	在	帕妥珠单抗	组	较为	常见	（由	霍夫曼–	罗氏...）。
	with	**placebo**	**group**	**compared,**	**diarrhea**	**in**	**pertuzumab**		**group**	**relatively**	**common**	**(by**	**F. Hoffman La Roche...)。**
(b) English	Diarrhea was more common with pertuzumab than with placebo. (Funded by F. Hoffmann–La Roche...).												

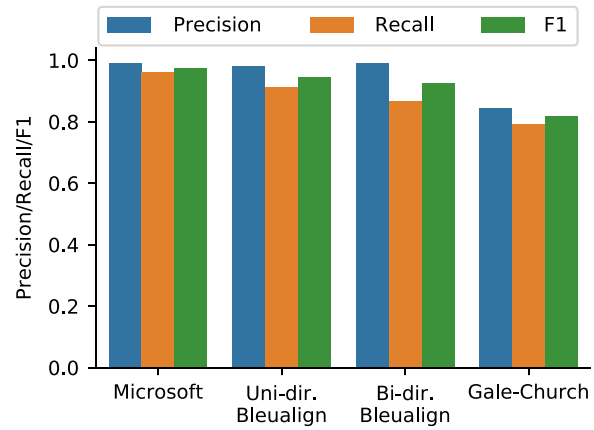

Fig. 6 Performance of three sentence aligners on the ParaMed corpus. Uni-directional Bleualign uses translations in zh→en only, whereas bi-directional Bleualign uses translations in both directions, giving it higher precision but lower recall

Table 6 Statistics of the ParaMed corpus

Language	Articles	Sentences	Avg. Len.	Tokens	Unique Tokens
English	1,966	97,441	31.08	3,028,434	55,673
Chinese			29.93	2,916,779	46,700

it should be emphasized that the ParaMed corpus is translated by human translators from English to Chinese and this bias will influence the machine translation quality [40].

To understand the translation quality as a function of in-domain dataset size, we fine-tuned the transformer model on 4,000, 8,000, 16,000, 32,000, 64,000 and all 93,303 sentence pairs (Fig. 7). For both zh→en and en→zh models, we saw improvement as the number of in-domain sentence pairs increased. The most significant improvement occurred at 4,000 sentence pairs (en→zh: +25.3 BLEU; zh→en: +13.4 BLEU). Translation quality continued to improve as the size of the dataset grows, albeit at a slower pace. Compared with baseline, the full dataset with 93,303 sentence pairs increased the BLEU score by 33.0 (24.3) points in en→zh (zh→en) directions. We observed similar effects on the LSTM model. The most significant improvement occurred at 4,000 sentence pairs (en→zh: +19.9 BLEU; zh→en: +13.7 BLEU). Compared with the baseline, the full ParaMed dataset increased the BLEU score by 28.1 (23.0) in en→zh (zh→en) directions.

To determine whether the pre-training on WMT18 data is necessary, we trained a *de novo* model using only ParaMed data, which was significantly faster than training a baseline model followed by fine-tuning. Compared with *de novo* training with a transformer model, pre-training on WMT18 baseline plus fine-tuning provided a meaningful boost in translation quality. Such boosts were most evident on small in-domain datasets. With 4,000 sentence pairs, pre-training improved the BLEU score by 34.1 (28.8) points for en→zh (zh→en) directions. The difference decreased as in-domain dataset grew, dropping to 7.9 (6.8) BLEU points for en→zh (zh→en) at the full-set level. A larger in-domain dataset is needed to completely compensate for the gap in translation quality. We observed similar effects for the LSTM model. The fine-tuned model consistently outperformed the *de novo* model for the size of the ParaMed dataset. The gap will likely become smaller as the dataset grows.

Machine translation error analysis

We showed two examples in this section to illustrate common mistakes made by our models. In the zh→en direction, the phrase "铂类-紫杉类" was correctly translated by the fine-tuned model to "platinum-taxane", and mistranslated by the baseline model to "Pt-Pseudophyllus" Table 7). The baseline model has not seen the phrase "紫杉类" during training and thus resulted in incorrect decoding. A similar situation occurred at the phrase "贝伐珠单抗". The fine-tuned model was able to correctly translate the phrase into bevacizumab, a chemotherapy medication, whereas the baseline model incorrectly decoded the phrase as "Bavaris mono-repellent".

Similar situations occurred for the en→zh direction (Table 8). Two medications, "olaparib" and "bevacizumab", were correctly translated by the fine-tuned model as "奥拉帕利" and "贝伐珠单抗", but incorrectly translated by the baseline model as "孤寡老人" and "白蜂". Fine-tuning on in-domain data extended the model

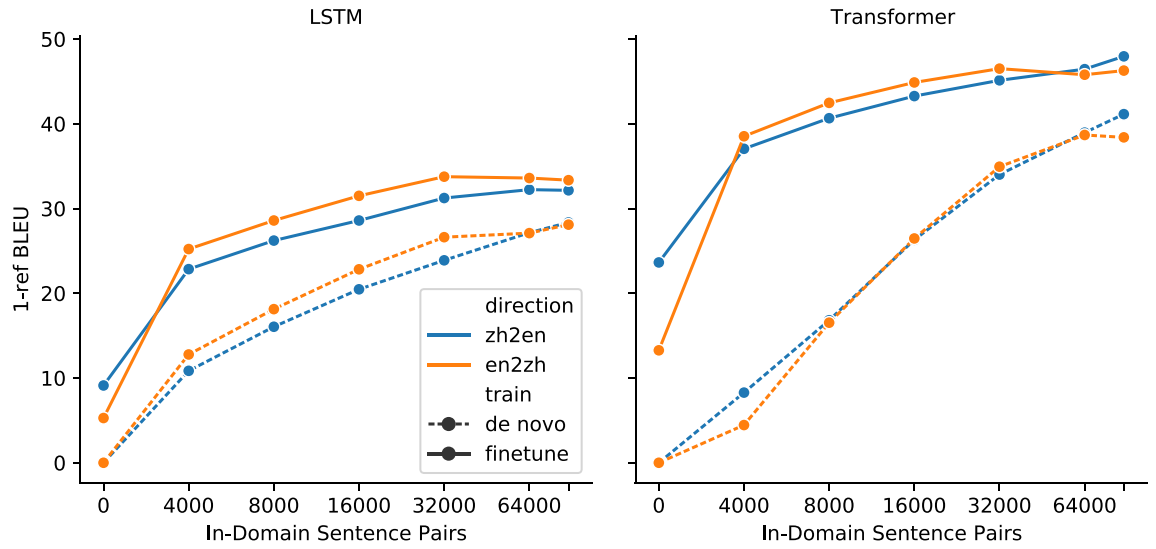

Fig. 7 Translation quality improves as the size of the ParaMed corpus increases. The solid lines correspond to the fine-tuned model, whereas the dashed lines correspond to the *de novo* model. The BLUE score for solid line at x = 0 shows the baseline model performance without fine-tuning

Table 7 铂类-紫杉类 and 贝伐珠单抗 were never seen by the baseline model and were translated incorrectly (red text)

(a) Source	*huànzhě*	*jiēshòu*	*bólèi*	*zǐshānlèi*	*yàowù*	*huàliáo*	*bèifázhūdānkàng*	*yīxiàn*	*zhìliáo*	*hòu*	
	患者	接受	铂类–	紫杉类	药物	化疗+	贝伐珠单抗	一线	治疗	后,	
	patients	**receiving**	**platinum**	**taxane**	**drug**	**chemotherapy**	**bevacizumab**	**first-line**	**treatment**	**afterwards**	
	běn	*yánjiū*	*yāoqiú*	*qí*	*bùnéng*	*yǒu*	*bìngbiàn*	*jìxiàng*	*huòzhě*	*zài*	*zhìliáo*
	本	研究	要求	其	不能	有	病变	迹象,	或者	在	治疗
	this	**study**	**require**	**they**	**don't**	**have**	**disease**	**evidence**	**or**	**after...**	**treatment**
	hòu	*dádào*	*línchuáng*	*wánquán*	*huò*	*bùfen*	*huǎnjiě*	*dìngyì*	*cānjiàn*	*biǎo*	
	后	达到	临床	完全	或	部分	缓解	（定义	参见	表1）。	
	...after	**achieve**	**clinical**	**complete**	**or**	**partial**	**relief**	**definition**	**see**	**table 1**	
(b) Target	After first-line treatment with platinum-taxane chemotherapy plus bevacizumab, patients were required to have no evidence of disease or to have had a clinical complete or partial response (definitions in Table 1).										
(c) NEJM translation	Patients were required to have no evidence of disease or to have a clinical complete or partial response after treatment after first-line platinum-taxane chemotherapy plus bevacizumab (as defined in Table 1).										
(d) WMT18 translation	after Pt-Pseudophyllus drug chemotherapy + Bavaris mono-repellent first-line treatment, the study required that the patient should not show signs of lesion or complete or partial clinical relief after treatment (see table 1 for definition).										

vocabulary and made it more accurate to decode medical terminology.

Conclusions

The popularity of neural machine translation models has boosted the need for large datasets. Public releases of many large-scale parallel corpora have significantly improved the quality of machine translation.

Machine translation in the biomedical domain has seen increasing attention in recent years [14, 41, 42]. Biomedical literature is rich in terminology for describing various diseases and biological processes. To add to this challenge, biomedical translation mandates a high standard of translation accuracy because the consequence of misinterpretation in medical decisions can be severe. All these challenges call for the curation of large-scale biomedical parallel corpora.

Despite the need for biomedical parallel text, curation of large-scale corpora have been biased towards European language pairs. Biomedical parallel corpora have been made available across several pairs of European languages, including English, German, Spanish, France, Portuguese, and Polish, to name a few. To our knowledge, there is no English-Chinese parallel corpus in the public domain.

Table 8 Olaparib and bevacizumab were not seen by the baseline model and were translated incorrectly (red text)

(a) Source	The lack of a maintenance olaparib monotherapy comparator group is a limitation of the PAOLA-1 trial, making it difficult to conclude whether the progression-free survival benefit seen in patients with HRD-positive tumors without BRCA mutations (who were not included in the SOLO1 trial) was due largely to the addition of olaparib or whether a synergistic effect occurred with olaparib and bevacizumab.														
(b) Target	*wèishèzhì*	*àolāpàlì*	*dānyào*	*wéichí*	*zhìliáo*	*dùizhàozǔ*	*shì*		*shìyàn*	*de*	*yígè*	*júxiànxìng*	*zhé*	*shǐdé*	
	未设置	奥拉帕利	单药	维持	治疗	对照组	是	PAOLA-1	试验	的	一个	局限性,	这	使得	
	lacking	**olaparib**	**monotherapy**	**maintenance**	**treatment**	**comparator**	**is**	**PAOLA-1**	**trial**	**'s**	**a**	**limitation**	**this**	**makes**	
	wǒmēn	*nányǐ*	*quèdìng*	*zài*	*wú*		*tūbiàn*	*de*		*yángxìng*	*zhǒngliú*	*huànzhě*		*shìyàn*	
	我们	难以	确定	在	无	BRCA	突变	的	HRD	阳性	肿瘤	患者	(SOLO1	试验	
	us	**difficult to**	**determine**	**in...**	**no**	**BRCA**	**mutation**	**'s**	**HRD**	**positive**	**tumor**	**patients**	**(SOLO1**	**trial**	
	wèi	*nàrù*	*cǐlèi*	*huànzhě*	*zhōng*	*guānchádào*	*wújìnzhǎn*	*shēngcúnqī*	*huòyì*	*de*	*yuányīn*	*shìfǒu*	*zhǔyàoshì*		
	未	纳入	此类	患者)	中	观察到	无进展	生存期	获益	的	原因	是否	主要是		
	not	**including**	**these**	**patients)**	**...in**	**observed**	**prog.-free**	**survival**	**benefit**	**'s**	**reason**	**whether**	**largely**		
	yóuyú	*jiāyòng*	*àolāpàlì*	*yě*	*wúfǎ*	*quèdìng*	*àolāpàlì*	*hé*	*bèifázhūdānkàng*	*shìfǒu*	*chǎnshēng*	*xiétóng*	*zuòyòng*		
	由于	加用	奥拉帕利,	也	无法	确定	奥拉帕利	和	贝伐珠单抗	是否	产生	协同	作用.		
	due to	**adding**	**olaparib**	**also**	**can't**	**determine**	**olaparib**	**and**	**bevacizumab**	**whether**	**produce**	**synergistic**	**effect.**		
(c) NEJM translation	*quēfá*	*àolāpàlì*	*wéichí*	*dānyào*	*zhìliáo*	*dùizhàozǔ*	*shì*		*shìyàn*	*de*	*yígè*	*júxiànxìng*	*yīncǐ*	*nányǐ*	
	缺乏	奥拉帕利	维持	单药	治疗	对照组	是	PAOLA-1	试验	的	一个	局限性,	因此	难以	
	lacking	**olaparib**	**maintenance**	**monotherapy**	**treatment**	**comparator**	**is**	**PAOLA-1**	**trial**	**'s**	**a**	**limitation,**	**therefore**	**difficult to**	
	quèdìng	*zài*	*wú*		*tūbiàn*	*de*		*yángxìng*	*zhǒngliú*	*huànzhě*	*wèi*	*nàrù*		*shìyàn*	
	确定	在	无	BRCA	突变	的	HRD	阳性	肿瘤	患者	(未	纳入	SOLO1	试验)	
	determine	**in**	**no**	**BRCA**	**mutation**	**'s**	**HRD**	**positive**	**tumor**	**patients**	**(not**	**included in**	**SOLO1**	**trial)**	
	zhōng	*guānchádàode*	*wújìnzhǎn*	*shēngcúnqī*	*huòyì*	*zhǔyào*	*shì*	*yóuyú*	*jiāyòngle*	*àolāpàlì*	*háishì*	*yóuyú*	*àolāpàlì*	*hé*	
	中	观察到的	无进展	生存期	获益	主要	是	由于	加用了	奥拉帕利,	还是	由于	奥拉帕利	和	
	in	**observed**	**prog.-free**	**survival**	**benefit**	**largely**	**is**	**due to**	**adding**	**olaparib,**	**or**	**due to**	**olaparib**	**and**	
	bèifázhūdānkàng	*chǎnshēngle*	*xiétóng*	*zuòyòng*											
	贝伐珠单抗	产生了	协同	作用.											
	bevacizumab	**produced**	**synergistic**	**effect.**											
(d) WMT18 translation	*quēfá*	*wéichí*	*gūguǎ*	*lǎorén*	*dānyī*	*liáofǎ*	*bǐjiào*	*xiǎozǔ*	*shì*		*shìyàn*	*de*	*yígè*	*xiánzhí*	
	缺乏	维持	孤寡	老人	单一	疗法	比较	小组	是	PAOLA-1	试验	的	一个	限制,	
	lacking	**maintenance**	**lonely**	**elderly**	**mono**	**therapy**	**comparator**	**group**	**is**	**PAOLA-1**	**trial**	**'s**	**a**	**limitation,**	
	shǐdé	*nányǐ*	*duàndìng*	*zài*	*méiyǒu*		*tūbiàn*	*de*		*yángxìng*	*zhǒngliú*	*huànzhě*	*wèi*	*bāokuò*	
	使得	难以	断定	在	没有	BRCA	突变	的	HRD	阳性	肿瘤	患者	(未	包括	
	making	**difficult to**	**determine**	**in...**	**no**	**BRCA**	**mutation**	**'s**	**HRD**	**positive**	**tumor**	**patients**	**(not**	**included**	
	zài		*zhìyànzhōng*	*zhōng*	*kàndàode*	*wújìnzhǎn*	*shēngcún*	*liyì*	*zhǔyào*	*shì*	*yóuyú*	*guǎtóu*	*yòuchóng*	*de*	
	在	SOLO1	试验中)	中	看到的	无进展	生存	利益	主要	是	由于	寡头	幼虫	的	
	in	**SOLO1**	**trial)**	**...in**	**observed**	**prog.-free**	**survival**	**benefit**	**largely**	**is**	**due to**	**oligo-**	**larva**	**'s**	
	zēngjiā	*háishì*	*yǔ*	*guǎtóu*	*hé*	*bái*	*fēng*	*fāshēng*	*xiétóng*	*zuòyòng.*					
	增加	还是	与	寡头	和	白	蜂	发生	协同	作用.					
	increase	**or**	**with**	**oligo-**	**and**	**white**	**bee**	**produce**	**synergistic**	**effect.**					

We have presented an English-Chinese parallel dataset in the biomedical domain. We have shown that a baseline model trained on out-of-domain data (WMT18) has limited generalizability to the biomedical domain and that as few as 4000 sentence pairs from the ParaMed dataset substantially improved translation quality. The translation quality continued to improve as the dataset grew. Further, pre-training with the out-of-domain data benefited translation quality, even at the full-set level.

We plan to expand our parallel corpus as New England Journal of Medicine continues to translate more articles. In the future, we would like to include bilingual PubMed abstracts as part of our parallel corpus.

Abbreviations

UMLS: Unified Medical Language System; WMT: Conference on Machine Translation; NEJM: The New England Journal of Medicine; SBD: Sentence boundary detection.

Acknowledgements

We thank Renjie Zheng, Baigong Zheng, Mingbo Ma, and Kenneth Church for their insights.

Authors' contributions

BL conceived the study and performed the analyses, BL and LH wrote the paper. Both authors have read and approved the manuscript.

Funding

We have no funding source to disclose.

Availability of data and materials

The code and data are available at https://github.com/boxiangliu/ParaMed.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Deep Learning, Baidu Research, Sunnyvale, USA. [2]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, USA.

Received: 22 October 2020 Accepted: 25 August 2021
Published online: 06 September 2021

References

1. Bamforth I. Biomedical translation. BMJ. 1998;316(7124):2–7124.
2. Das A. Medical interpreters. BMJ. 2009;338:2354.
3. Hassan H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C, Huang X, Junczys-Dowmunt M, Lewis W, Li M, et al. Achieving human parity on automatic chinese to english news translation; 2018. arXiv preprint arXiv:1803.05567
4. Bodenreider O. The unified medical language system (UMLs): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl–1):267–70.

5. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data; 2015. arXiv preprint arXiv:1511.06709
6. Duh K, Neubig G, Sudoh K, Tsukada H. Adaptation data selection using neural language models: experiments in machine translation. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2013. p. 678–683.
7. Luong M-T, Manning CD. Stanford neural machine translation systems for spoken language domains. In: Proceedings of the international workshop on spoken language translation; 2015. p. 76–79.
8. Neves M. A parallel collection of clinical trials in Portuguese and English. In: Proceedings of the 10th workshop on building and using comparable corpora; 2017. p. 36–40
9. Bawden R, Di Nunzio G, Grozea C, Unanue I, Yepes A, Mah N, Martinez D, Névéol A, Neves M, Oronoz M, et al. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In: 5th conference on machine translation; 2020.
10. Dušek O, Hajič J, Hlaváčová J, Libovický J, Pecina P, Tamchyna A, Urešová Z. Khresmoi Summary Translation Test Data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University; 2017. http://hdl.handle.net/11234/1-2122
11. Villegas M, Intxaurrondo A, Gonzalez-Agirre A, Marimon M, Krallinger M. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In: Malero M, Krallinger M, Gonzalez-Agirre A, editors. LREC MultilingualBIO: Multilingual Biomedical Text Processing. 2018.
12. Tiedemann J. Parallel data, tools and interfaces in opus. LREC. 2012;2012:2214–8.
13. Tian L, Wong DF, Chao LS, Quaresma P, Oliveira F, Yi L. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In: LREC; 2014. p. 1837–1842
14. Barrault L, Bojar O, Costa-jussà MR, Federmann C, Fishel M, Graham Y, Haddow B, Huck M, Koehn P, Malmasi S, et al. Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the fourth conference on machine translation (Volume 2: Shared Task Papers, Day 1); 2019. p. 1–61.
15. Buck C, Koehn P. Findings of the wmt 2016 bilingual document alignment shared task. In: Proceedings of the first conference on machine translation: Volume 2, Shared Task Papers; 2016. p. 554–563.
16. Gomes L, Lopes GP. First steps towards coverage-based document alignment. In: Proceedings of the first conference on machine translation: volume 2, Shared Task Papers; 2016. p. 697–702.
17. Read J, Dridan R, Oepen S, Solberg LJ. Sentence boundary detection: a long solved problem? In: Proceedings of COLING 2012: Posters; 2012. p. 985–994.
18. Alias-i: Alias-i. http://alias-i.com/lingpipe. Accessed:2019-12-10 (2008)
19. Bird S, Loper E, Klein E. Natural language processing with Python. Newton: O'Reilly Media Inc.; 2009.
20. Gale WA, Church KW. A program for aligning sentences in bilingual corpora. Comput Linguist. 1993;19(1):75–102.
21. Moore RC. Fast and accurate sentence alignment of bilingual corpora. In: Conference of the association for machine translation in the Americas, Springer; 2002 p. 135–144.
22. Varga D, Halácsy P, Kornai A, Nagy V, Németh L, Trón V. Parallel corpora for medium density languages. Amsterdam studies in the theory and history of linguistic science series 4. 2007;292:247.
23. Ma X. Champollion: a robust parallel text sentence aligner. In: LREC; 2006. p. 489–492.
24. Sennrich R, Volk M. Iterative, MT-based sentence alignment of parallel texts. In: Proceedings of the 18th Nordic conference of computational linguistics (NODALIDA 2011); 2011. p. 175–182.
25. Sennrich R, Volk M. Mt-based sentence alignment for ocr-generated parallel texts. In: The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010); 2010.
26. Koehn P, Khayrallah H, Heafield K, Forcada ML. Findings of the WMT 2018 shared task on parallel corpus filtering. In: Proceedings of the third conference on machine translation: shared task papers; 2018. p. 726–739.
27. Junczys-Dowmunt M. Dual conditional cross-entropy filtering of noisy parallel corpora; 2018. arXiv preprint arXiv:1809.00197
28. Muthukadan B. Selenium with Python. https://selenium-python.readthedocs.io/. Accessed 10 Dec 2019
29. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al. Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion Volume proceedings of the demo and poster sessions; 2007. p. 177–180.
30. Junczys-Dowmunt M. eserix. https://github.com/emjotde/eserix. Accessed 10 Dec 2019
31. Ziemski M, Junczys-Dowmunt M, Pouliquen B. The united nations parallel corpus v1. 0. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16); 2016. p. 3530–3534.
32. Ramírez-Sánchez G, Zaragoza-Bernabeu J, Bañón M, Ortiz-Rojas S. Bifixer and bicleaner: two open-source tools to clean your parallel data. In: Proceedings of the 22nd annual conference of the European Association for Machine Translation; 2020. p. 291–298. European Association for Machine Translation, Lisboa, Portugal.
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser, Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.
34. Klein G, Kim Y, Deng Y, Senellart J, Rush A. OpenNMT: Open-source toolkit for neural machine translation. In: Proceedings of ACL 2017, system demonstrations; 2017. p. 67–72. Association for Computational Linguistics, Vancouver, Canada. https://www.aclweb.org/anthology/P17-4012
35. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
36. Bojar O, Federmann C, Fishel M, Graham Y, Haddow B, Huck M, Koehn P, Monz C. Findings of the 2018 conference on machine translation (wmt18). In: Proceedings of the third conference on machine translation, Volume 2: Shared Task Papers; 2018. p. 272–307. Association for Computational Linguistics, Belgium, Brussels. http://www.aclweb.org/anthology/W18-6401
37. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units; 2015. arXiv preprint arXiv:1508.07909
38. Kingma DP, Ba J. Adam: a method for stochastic optimization; 2014. arXiv preprint arXiv:1412.6980.
39. Sánchez-Cartagena VM, Bañón M, Rojas SO, Ramírez-Sánchez G. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In: Proceedings of the third conference on machine translation: shared task papers; 2018. p. 955–962.
40. Graham Y, Haddow B, Koehn P. Translationese in machine translation evaluation; 2019. arXiv preprint arXiv:1906.09833
41. Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huck M, Yepes AJ, Koehn P, Logacheva V, Monz C, et al. Findings of the 2016 conference on machine translation. In: Proceedings of the first conference on machine translation: Volume 2, Shared Task Papers; 2016. p. 131–198.
42. Bojar O, Chatterjee R, Christian F, Yvette G, Barry H, Matthias H, Philipp K, Qun L, Varvara L, Christof M, et al. Findings of the 2017 conference on machine translation (wmt17). In: Second Conference onMachine Translation; 2017. p. 169–214. The Association for Computational Linguistics

Publisher's Note