Using Bilingual Patents for Translation Training

John Lee, Benjamin Tsou, Tianyuan Cai
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR
{jsylee, rlbtsou}@cityu.edu.hk, tianycai-c@my.cityu.edu.hk

Abstract

While bilingual corpora have been instrumental for machine translation, their utility for training translators has been less explored. We investigate the use of bilingual corpora as pedagogical tools for translation in the technical domain. In a user study, novice translators revised Chinese translations of English patents through bilingual concordancing. Results show that concordancing with an in-domain bilingual corpus can yield greater improvement in translation quality of technical terms than a general-domain bilingual corpus.

1 Introduction

Text corpora are increasingly used in language pedagogy, with many studies showing their effectiveness in data-driven language learning (Boulton, 2017), language exercise generation (Susanti et al., 2018) and assisted writing (Chang and Chang, 2015), among many other tasks. Bilingual corpora, however, remain underused in translation pedagogy. Most research has focused on exploiting bilingual data for training machine translation (MT) systems and for developing translation memory (TM) in computer-assisted translation tools.

Professional translators generally favor dictionaries and search engines over corpora (Man et al., 2020). While these tools may be sufficient when translators work in a familiar domain, more domain-specific examples are often needed to handle unfamiliar terms and collocations in specialized areas. Complementing MT and TM, bilingual corpora allow translators to consult existing examples in context to evaluate the appropriateness of a candidate translation (Bowker and Barlow, 2008). Even so, many translators never used corpora (Gallego-Hernández, 2015; Gough, 2013). Translation students were more familiar with MT and TM than corpora, which are accessible only to a minority (Man et al., 2020). It is therefore worthwhile to evaluate the potential of this underused resource as pedagogical tool.

2 Research Questions

This paper investigates the use of bilingual corpora for translation training in the specialized domain of technical writing, where the preponderance of rare terms and collocations makes translation aid critical. Our study focuses on patents, whose translation market has been rapidly expanding as patents are filed in a greater variety of languages across the globe. We aim to address two research questions:

Self-learning via concordancing: Are novice translators capable of improving translation drafts via independent concordancing with bilingual corpora?

General-domain vs. in-domain bilingual data: Are in-domain bilingual corpora more effective than general-domain corpora as pedagogical tools for translation?

The rest of the paper is organized as follows. After a summary of previous work (Section 3), we describe the corpus data in our study (Section 4). We then present the design of our study, in which

Proceedings of the 28th International Conference on Computational Linguistics, pages 3461–3466
Barcelona, Spain (Online), December 8-13, 2020

novice translators produced drafts with MT outputs and subsequently revised them via concordancing with general-domain and in-domain parallel corpora (Section 5). Finally, we report results (Section 6) and conclude (Section 7).

3 Previous work

Corpus use often requires proficiency with query languages and other technical skills. Despite the learning curve, previous research suggests that corpora can be effective as pedagogical tools for language learning. Wang (2001) demonstrated students' ability to induce different meanings of a Chinese word with a Chinese-English bilingual concordancer. Gao (2011) showed that L2 learners can refine their writing, including word choices and word combinations, via a parallel concordancer without guidance from instructors. In a study by Gaskell and Cobb (2004), up to 70% of the students were able to make corrections based on concordance feedback. A substantial number of the students became independent, persistent corpus users.

Bilingual corpora have been increasingly used in a variety of translation studies (Prieto Ramos, 2019). Research on corpus-driven translation training has investigated specialized domains such as insurance and legal texts (Monzó Nebot, 2008). Pedagogical benefits have been demonstrated in using bilingual corpora for English-Spanish translation of insurance documents (Corpas Pastor and Seghiri, 2009). An online legal and documentary bilingual corpus has been shown to enhance self-learning of legal text (Fan and Xu, 2007). In a Master's program in legal translation, students compiled a monolingual English corpus as an aid for Spanish-to-English translation of judgments (Ramos and Moreno, 2016). To the best of our knowledge, no evaluation has been reported for corpus-based translation training in the domain of technical writing, likely due to a lack of bilingual data. This bottleneck has been eased with the compilation of PatentLex (Lu et al., 2009), a large-scale collection of comparable Chinese-English patents which will be used in our study.

4 Data

General-domain corpus OPUS2, a large collection of freely available multilingual data (Tiedemann, 2009), served as the general-domain parallel corpus in our study. It includes text from political and administrative sources, user-provided movie subtitles, as well as software localisation, news providers, translated descriptions of medical products, religious texts and multilingual wikis and other websites.

In-domain corpus Composed of over 300K Chinese and English patents, PatentLex (Tsou et al., 2019) has served as the dataset for patent MT shared tasks such as those organized by NTCIR workshops. Our in-domain bilingual data consisted of over 30 million parallel bilingual sentences extracted from PatentLex.

5 Evaluation Set-up

5.1 Subjects

Our subjects were 31 students in a postgraduate course in Linguistics at City University of Hong Kong. All students were native speakers of Chinese who were proficient in English, having met the university's admissions requirement for IELTS. Our study focused on English-to-Chinese translation. This translation direction avoids L2-related errors in the target text and thus serves to highlight genuine translation difficulties.

5.2 Translation procedure

We selected 10 English patent abstracts from PatentLex, and used their manually translated Chinese versions as gold data. The average length was 115.2 words for the English abstracts and 206.6 characters for the Chinese versions. Each student was randomly assigned one of these ten abstracts, and asked to submit three translation versions, as follows:

Baseline version The student first produced a translation draft on the basis of the output of Google Translate (http://translate.google.com) and Baidu (http://fanyi.baidu.com). Even though many technical terms in the abstract may be unknown to the student, no additional assistance was allowed.

General-domain version The student revised the baseline version via parallel concordancing, consulting corpus examples in OPUS2 (Section 4) through Sketch Engine (Kilgarriff et al., 2008). As a general-domain corpus, OPUS2 has more limited coverage of technical terms, and their translation may not necessarily reflect the terminology adopted in patents.

In-domain version The student produced a final version via parallel concordancing with in-domain examples in PatentLex (Section 4). The search interface for PatentLex supported bilingual keyword search in the KWIC format.[1]

5.3 Translation quality assessment

We evaluated the quality of the above translation versions both automatically and manually. In automatic evaluation, we followed Li et al. (2012) in using the BLEU score (Papineni et al., 2002). Manual evaluation focused on "technical terms", broadly defined as noun phrases that bear a scientific or technical concept. Three human judges, all native speakers of Chinese with a bachelor's or master's degree in linguistics or language studies, performed the evaluation.

Gold data. One of the judges identified all technical terms that appeared in the ten English abstracts, and then aligned them to their counterpart in the gold Chinese translations. After review by the other two judges, the list consisted of a total of 200 technical terms, amounting to an average of 20 terms per abstract.

Quality assessment. The three judges independently identified the translation of each technical term in the MT output from Google Translate and Baidu, and in the three translation versions by the students (Section 5.2). The judges then assessed the translation quality of each term. A translated term was labeled as "equivalent" if it was identical to the gold, or constituted an acceptable alternative; otherwise, it was labeled as "worse".

We calculated pairwise agreement among the three annotators on their labeling of the quality of the translated terms, excluding those that were identical to the gold. The kappa values are 0.4654, 0.4757, and 0.5078, respectively, all at the level of "moderate agreement" (Landis and Koch, 1977). The final label for each technical term was determined by majority vote.

6 Results

Table 1 reports the evaluation results, providing both the overall score and a breakdown into individual abstracts, each of which was assigned to 2 to 4 students.

Comparison among student versions. We first compare the translation accuracy of technical terms in the three student versions (Section 5.2). The baseline version attained a higher accuacy (63.25%) than the general-domain version (62.80%). The slight decrease in accuracy suggests possible translation divergences between the general-domain data and the specialized domain of technical writing. In contrast, in-domain concordancing helped increase the accuracy to 65.12%, above the baseline. The in-domain version (0.2430) also outperformed the baseline (0.2290) in terms of BLEU score. These results show that in-domain bilingual data can help novice translators revise the translation of technical terms, and can be more helpful than general-domain bilingual data.

Among the ten abstracts in the evaluation dataset, eight achieved higher accuracy in their in-domain version than in their general-domain counterpart. Six of the ten abstracts yielded in-domain versions that outperformed the baseline version.

MT vs. student versions. The MT outputs from Google and Baidu were more accurate in the translation of technical terms (66.25% and 69.65%, respectively) than the student versions, even after in-domain concordancing (65.12%). The BLEU score reveals a similar gap, with Google output (0.2519) and Baidu

[1] Accessed at http://patentlex.chilin.hk in April 2020. Content from the ten selected abstracts was not available through the search interface.

Translation version →		MT versions		Student versions		
↓ Metric		Google output	Baidu output	Baseline version	General-domain version	In-domain version
BLEU	Average	0.2519	0.2614	0.2290	0.2207	**0.2430**
Technical term translation accuracy	Average	66.25%	69.65%	63.25%	62.80%	**65.12%**
	Individual abstracts (10)	10.53%	15.79%	**14.29%***	12.24%*	14.00%*
		60.71%	60.71%	62.50%*	62.03%*	**67.09%***
		66.67%	73.33%	63.33%	66.67%	**71.43%***
		67.74%	72.41%	67.19%	66.13%	**70.97%***
		100.00%	95.65%	95.38%	**100.00%**	98.46%*
		50.00%	55.56%	29.63%	29.63%	**33.33%**
		69.23%	69.23%	63.46%	63.46%	**65.38%**
		64.71%	76.47%	**65.93%***	62.96%	62.50%
		81.25%	85.71%	**79.17%**	78.72%	**79.17%**
		91.67%	91.67%	**91.67%**	86.11%	88.89%

Table 1: BLEU score and translation accuracy of technical term. Translation accuracy is also shown for individual abstracts, with an asterisk placed on student versions that are more accurate than at least one of the MT versions.

output (0.2614) both outperforming the in-domain version (0.2430). The lower scores likely reflected the fact that the subjects were neither experienced translators nor specialists in technical writing.

The goal of this study is not to show that PatentLex can help human translators outperform MT systems. Human performance depends on a range of factors, including the translators' competence and their background in technical writing, as well as the patent content. As it turned out, for 5 of the 10 abstracts (marked with asterisks in Table 1), the novice translators produced an in-domain version that was more accurate than at least one of the MT versions. In future work, we plan to investigate whether semantic similarity scores can predict the reliability of corpus examples in translating a term.

Translation difficulties. We analyze the words that were most prone to mistranslation by the students in the in-domain versions. Among terms attested in the bilingual data, those with multiple translation alternatives (e.g., "module") tend to be more difficult than those with fewer alternatives (e.g., "tank"). Consider the term "first tank", for example in the sentence "The frequency generating means has first tank means for resonating at the first frequency." Translated by Google into the term 第一谐振器装置 *diyi xiezhenqi zhuangzhi*, it was revised by three out of four students to match the gold standard 第一槽路装置 *diyi caolu zhuangzhi* 'first tank'. In contrast, consider the term "module", for example in the sentence "A packaging machine having multiple modules each performing a separate function in a packaging sequence". Both MT systems translated it into 模块 *mokuai* 'module', but no student discerned the most appropriate term 标准单元 *biaozhun danyuan* 'module' from the example bilingual sentences in PatentLex.

7 Conclusions

We have presented the first evaluation on the use of bilingual corpora for translation training in technical writing. In a user study, novice translators translated English patent abstracts into Chinese with assistance from bilingual examples in general-domain and in-domain corpora. Results suggest that they were able to exploit bilngual data to improve translation quality of technical terms. An in-domain corpus of patents appears to be more effective as pedagogical tool than a general-domain corpus, with respect to both term translation accuracy and BLEU score.

Acknowledgements

This work was partially supported by CityU Internal Funds for ITF Projects (no. 9678104).

References

Alex Boulton. 2017. Data-driven Learning and Language Pedagogy. In S. Thorne and S. May, editors, *Language, Education and Technology: Encyclopedia of Language and Education*. Springer, New York, NY, USA.

Lynne Bowker and Michael Barlow. 2008. A comparative evaluation of bilingual concordancers and translation memory systems. In Elia Yuste Rodrigo, editor, *Topics in Language Resources for Translation and Localisation*, pages 1–22. John Benjamins, Philadelphia, PA.

Jim Chang and Jason S. Chang. 2015. WriteAhead2: Mining Lexical Grammar Patterns for Assisted Writing. In *Proc. NAACL-HLT*, page 106–110.

Gloria Corpas Pastor and Miriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English–Spanish). In A. Beeby, P. Rodríguez-Inés, and P. Sánchez-Gijón, editors, *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate*, page 75–107. John Benjamins.

May Fan and Xunfeng Xu. 2007. An evaluation of an online bilingual corpus for the self-learning of legal english. *System*, 30(1):47–63.

Daniel Gallego-Hernández. 2015. The use of corpora as translation resources: a study based on a survey of Spanish professional translators. *Perspectives: Studies in Translatology*, 23(3):375–391.

Zhao-Ming Gao. 2011. Exploring the effects and use of a chinese–english parallel concordancer. *Computer Assisted Language Learning*, 24(3):255–275.

Delian Gaskell and Thomas Cobb. 2004. Can Learners Use Concordance Feedback for Writing Errors? *System*, 32:301–319.

Joanna Gough. 2013. *Survey of professional translators' use of on-line resources for terminology research*. Ph.D. thesis, University of Surrey.

Adam Kilgarriff, Mils Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proc. EURALEX*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.

Xianhua Li, Yao Meng, and Hao Yu. 2012. Improving chinese-to-japanese patent translation using english as pivot language. In *Proc. 26th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 117–126.

Bin Lu, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Oi Yee Kwong. 2009. The Construction of a Chinese-English Patent Parallel Corpus. In *Proc. MT Summit XII: Third Workshop on Patent Translation*, pages 17–24.

Deliang Man, Aiping Mo, Meng Huat Chau, John Mitchell O'Toole, and Charity Lee. 2020. Translation technology adoption: evidence from a postgraduate programme for student translators in China. *Perspectives*, 28(2):253–270.

Esther Monzó Nebot. 2008. Corpus-based activities in legal translator training. *Interpreter and Translator Trainer*, 2(2):221–251.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*.

Fernando Prieto Ramos. 2019. The use of corpora in legal and institutional translation studies: Directions and applications. *Translation Spaces*, 8(1):1–11.

María Del Mar Sánchez Ramos and Francisco J. Vigier Moreno. 2016. Using monolingual virtual corpora in public service legal translator training. In Elena Martin-Monje, Izaskun Elorza, and Blanca García Riaza, editors, *Technology-Enhanced Language Learning for Specialized Domains: Practical applications and mobility*.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic Distractor Generation for Multiple-choice English Vocabulary Questions. *Research and Practice in Technology Enhanced Learning*, 13(15).

Jörg Tiedemann. 2009. News from OPUS: a Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing*, 5:237–248.

Benjamin K. Tsou, Kapo Chow, Junru Nie, and Yuan Yuan. 2019. Towards a Proactive MWE Terminological Platform for Cross-Lingual Mediation in the Age of Big Data. In *Proc. 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 116–121.

Lixun Wang. 2001. Exploring Parallel Concordancing in English and Chinese. *Language Learning and Technology*, 5(3):174–184.