Translating Terminologies: A Comparative Examination of NMT and PBSMT Systems

Long-Huei Chen
The University of Tokyo
Tokyo, Japan
longhuei@g.ecc.u-tokyo.ac.jp

Kyo Kageura
The University of Tokyo
Tokyo, Japan
kyo@p.u-tokyo.ac.jp

Abstract

Terminology translation is a critical aspect in translation quality assurance, as it requires exact forms not typically expected of conventional translation. Recent studies have examined the quality of machine translation, but little work has focused specifically on the translation of terms. We present a comparative evaluation of the success of NMT and PBSMT systems in term translation. We selected eight language pairs among English, French, German, Finnish, and Romanian, taking into account their diverse language families and resource abundance. Based on the evaluation of Exact Match (EM) and recall scores, we concluded that NMT, in general, performs better with context, but PBSMT outperforms when translating without context, and found that significant differences often arise from language nature.

1 Introduction

Term translation is an important facet of translation quality assurance. Since terminologies are essential for communication among domain experts, term forms need to be consistent and context-independent to maintain the integrity of the underlying conceptual system during knowledge exchange (Sager, 1990). As such, term banks (collections of cross-lingual, cross-domain terminologies) ensure correct term usage across languages in the translation pipeline of humans.

The rise of machine learning in recent years has, for better or for worse, changed the landscape of translation forever. The typical evaluation of machine translation, due to a requirement of fast, automatic metrics during the training phase, typically involves the comparison with a set of human translation in what is calculated as the BLEU or the NIST scores of the translation (Papineni et al., 2002; Doddington, 2002). These approaches run counter to widely accepted frameworks of translation quality assurance (Görög, 2014; Peter et al., 2016) as the measures do not single out aspects of translation that humans traditionally attach importance.

Machine translation in general does not produce the exactness in forms required in term translation. Unlike translation of a text, where target text similar in meanings are equivalent as long as they fulfill the required functions, translated term forms must adhere to term banks (Kageura and Marshman, 2019). Machine translation also has implications in terminology building during the human translation process, as it can provide an automatic way to generate and validate the terminology resource that is available to translators (House, 2014; Chiocchetti and Lusicky, 2017; Yamada and Onishi, 2019). This is why we are also interested in learning how well the machine translation systems perform in term translation without context (Matis, 2010).

Here we present a comparative evaluation in the effectiveness of machine translation for terminology transfer across multiple languages. We investigate language pairs of varying training resource abundance on different machine translation architecture to understand the underlying factors of the effectiveness of terminology translation. We test systems with bidirectional translations and validate the terminology equivalence by referring to an established term bank.

2 Related Work

The Special Case for Terminology Translation

Traditional translators often approach terminology translation within the lenses of semasiological assumptions and treat terminology as a type of lexical elements (Achkasov, 2014). Nevertheless, when we take on an onomasiological point of view and understand that terms (Adamska-Salaciak, 2010; Lyding et al., 2006) are essentially definitions of concepts, then the degrees of equivalence are expected to be higher.

A key aspect of terminology translation is that the formation of a term in some language/domain is not solely at the discretion of the translator, but has structural, pragmatic, functional, and stylistic aspects that need to be taken into account (Achkasov, 2014). This produces a need for translation of terminologies that takes into account the domain terminology that is in existence (Kageura, 2012; Leitchik and Shelov, 2007)

Terminology in Translation Quality Assurance

Accuracy of terminologies in translated work is an essential element in translation quality assessment (Arango-Keeth and Koby, 2003; Görög, 2014; Peter et al., 2016). According to the standards establishing the essentiality of special treatments of terminology translation (ISO, 2010), policies relating to the adaptation of terminologies in translation work is necessitated and needs to be widely implemented. Substantial efforts have been expended in the past to evaluate the state of terminology translation in both phrase-based statistical machine translation (PBSMT) and more recently neural machine translation (NMT).

Yin et al. (2013) investigated consistency of terminology translation by cross-referencing patent documents in English and Chinese. Vintar (2018) evaluated both PBSMT and NMT between English and Slovene, a relatively lower-resource language pair. She concluded that thought Google's NMT serves a large amount of user-generated content at a large scale, the accuracy of its terminology translation within text leaves something to be desired.

Empirical Evaluation of NMT/SMT with Textual Corpora

Several studies examined the effectiveness of neural machine translation and statistical machine translation when applied to general text. Wu et al. in their original paper describing Google's NMT system (2016) observed increased performance compared to their previous public PBSMT system. Shterionov et al. (2017) conducted a comprehensive study on text translation and found NMT improved performances in multiple metrics as evaluated by humans. Dowling et al. (2018) tested both NMT and SMT systems on a lower-resourced language that is Irish and found that a domain-specific SMT system in some cases outperform NMT.

Muzaffar and Behera (2018), on the other hand, examined translation results in English-Urdu, a relatively resource-poor language pair, and concluded that NMT brings forward better comprehensibility and grammaticality. Castilho et al. (2017) recruited professional translators and found that as a tool for translators, NMT results do not reduce post-editing time compared to PBSMT. Work on specific genre includes (Toral and Way, 2018), which examine NMT vs. PBSMT performances on literary work, and found that NMT significantly increased the readability of the text for human readers. Kinoshita et al. (2017) examined the use of NMT and SMT in the translation of patent documents and concluded that NMT is superior in terms of human evaluations.

3 Machine Translation Models

3.1 Neural Machine Translation (NMT)

3.1.1 The Encoder-decoder Architecture

The basic structure of the modern neural machine translation system involves the encoding of a series of source text tokens, which can be words or sub-word unit encoding, into a hidden state representation (Cho et al., 2014).

$$z = \text{ENCODE}\left(w^s\right)$$

$$w^t|w^s \sim \text{DECODE}(z)$$

where z is the learned hidden state, w refers to the distributional representation of words, with the suffix s or t referring to source or target origin. In the simplest sequence-to-sequence architecture, the encoder hidden state came is learned from the long short-term memory (LSTM) unit applied on the source sentence words (Sutskever et al., 2014).

The encoder hidden state is then passed along to the decoder, which is then passed along to the decoder for output. The decoder generates the target sentence token-by-token while continuous updating its internal state. In addition, neural attention mechanisms encourage compositional decoding by taking into account the context in the decoder. At each step in the decoding, an attentional score is calculated from the decoding hidden state along with the encoding sentence tokens.

3.1.2 Zero-shot Translation

As Google's Neural Machine Translation system takes input from any training pairs across all languages (Johnson et al., 2017), cross-lingual transfer learning was made possible with the addition of a language-specific token designating the output languages. The same shared parameters are applied to allow for translation into any target languages. As a result, even when parallel data are lacking across specific language pairs, resulting in the so-called zero-shot translation, which is impossible in previous systems. This allows the highly-effective use and wide coverage of Google's systems, even in cases where parallel corporal resources are lacking for specific language pairs.

3.2 Phrase-Based Statistical Machine Translation (PBSMT)

Statistical machine translation models statistically enumerating and maximizing the adequacy and fluency of the target translation by maximizing the probability across all possible assignments, usually with the expectation-maximization (EM) algorithm. Phrase-based statistical machine translation (PBSMT) extends this approach to account for the fact that phrases often form the smallest unit of translation, and allows for phrase-level alignments to suggest the most likable translation.

4 Approach

We conduct our experiment by pairing a term bank, which are sources of cross-lingual translations of specialized terms, with a set of technical documents in which the translators are expected to adhere to the term source. We extract sentence pairs from the documents by searching for a context where the term appears in accordance with the term bank.

4.1 Data Source

The **Inter-Active Terminology for Europe (IATE)** (Johnson and Macphail, 2000) is the official term bank sanctioned by the European Union (EU). It is the go-to source with approximately 1.4 million multilingual entries of terminologies containing the cross-lingual translation of terms for translators working with the official European Union languages. The **European Parliament Proceedings (EuroParl) parallel corpus** is extracted from the proceedings of the European Parliament and includes versions in 24 European languages (Koehn, 2005). Size of the parallel corpora differs across language pairs, ranging from 400,000 to around 2.2 million sentence pairs.

Since IATE is the official EU-wide terminology as maintained and consulted by the translators under EU's employment, the combination of the two reflects the typical translation procedure when a commonly-agreed term source is provided for translators.

Source	Target	Sentence Pairs
English	French	58362
French	English	53470
English	German	38879
German	English	38879
English	Finnish	30994
Finnish	English	17486
English	Romanian	7676
Romanian	English	5151

Table 1: Size of the source-term sentence pairs, where only the source sentence is validated to contain the source term while the target term may or may not contain the target term.

Source/Target	Sentence Pairs
English-French	21057
English-German	14070
English-Finnish	17486
English-Romanian	2685

Table 2: Size of the human-validated sentence pairs, where the sentence pair is validated such that both source/target sentences contain the source/target term translation.

4.2 Language Pairs and Data Size

We choose to investigate four language pairs of the EuroParl parallel corpus, namely English-French (en-fr), English-German (en-de), English-Finnish (en-fi) and English-Romanian (en-ro). The languages are chosen by taking into account language families and data sizes.

1. **Source-term sentence pairs** are extracted from the corpora, and the source sentence is guaranteed to contain the source term, but the human-translated target sentence may or may not contain the target term. (Table 1)

2. **Human-validated term sentence pairs** are

those where both the source and the target sentences contain the source/target term from the terminology. They can be regarded as cases that the context of the sentences is guaranteed to reflect the definition of the terminology (Table 2)

We consider the second dataset as an equivalent operation when the human-translator chose the exact term translation as it appears in the term bank. This replaces the needs for human terminologists manually annotating the dataset, as the term bank has already been validated.

4.3 Google Cloud Translation APIs

The Google Cloud Translation API provides a programmatic interface for translating sentences across the supported languages using state-of-the-art translation models. The APIs include two models, the "nmt" model which is their new NMT model, and a "base" model, which as stated is a PBSMT model. We query the APIs to apply the model as needed.

4.4 Evaluation: Exact Match (EM) and Recall Scores

We compare term occurrence in results coming from target text produced from Google's Translation APIs and those from the official, human-translated target text. We presume that in cases where the term bank entry is present in the human-translated or machine-translated sentences, the term use in these cases are validated and considered correct usage.

Rather than using traditional measures of translation quality, in this work, we are mainly concerned with the success of different systems in their adequate reproduction of the relevant terminologies in the target text. Specifically:

1. **Exact Match (EM)** scores is defined as the exact occurrence of the ground truth target terms in the translated target sentence.

2. **Recall** is defined as the fraction of known target term words that occur in the target text.

For our evaluation, we do not make a distinction between the infections of terms. We chose this strict interpretation of exact match as we want to see how well these machine translation systems can fare in creating terminology resources for translators without context, in which case the exact form (including inflections) must be properly transferred across language barriers. The same scheme is also applied for with context translation

We recognize that, since both MT systems and human translation do not include an annotation as to the exact location of the term translation in the sentence, we are unable to verify the precision of the term translation or the F1 score. Also, we argue that since terms, unlike most multi-worded expressions, are technical in nature and have specific forms, it is less likely to occur by random in the target sentence and not as a translation, justifying our automated approach to evaluation.

5 Experiments

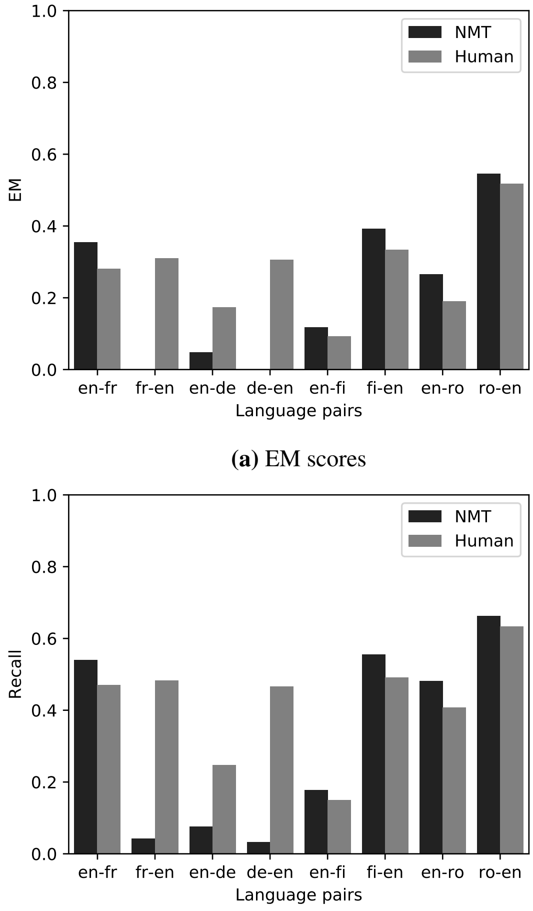
5.1 Adherence to Term Banks: Human v. NMT

In the first experiment, we apply translation systems to the source-term sentence pairs as detailed in §4.2. We compare performances of the system with the human translated sentences on how much the term bank target term is correctly translated in the target sentence. Results are given in Figure 1.

Cases in which the NMT scores are higher than that of human translations should *not* be interpreted as NMT performing better, but that the NMT systems adhere more to the term bank in a rote way. Humans may make the call on whether the particular term entry is applicable, or choose to use pronouns to avoid repetition of terms, and our evaluation may exclude the term variation deemed acceptable by humans.

- Despite varying human performances, NMT surpasses human scores for some language pairs but performs worse than human in others. This reflects the discongruity in a single end-to-end language model in its treatment of language pairs (§3.1.2).

- We observe that languages where parallel corpora resources are plentiful achieve lower NMT scores compared to the human scores. This suggests more parallel training data may shift the model's focus to language modeling and fluency rather than simple phrase-level correspondence, and indirectly hurt performance (§3.1).



(a) EM scores

(b) Recall scores

Figure 1: Comparing human performances with NMT, using datasets where only the source sentence with context is validated to contain the source term. We test the recall and EM scores with regards to the term bank translation and the target sentence.

5.2 Translating Term with and without Context: NMT v. PBSMT

In these experiments, we apply the human-validated sentence pairs to the MT systems; this set includes both source/target terms in their source/target sentences, so the terms are human-validated to reflect the context.

For resulting scores in Figure 2a and 2c, we translate the terms along with the context sentences and observe how well the translated sentences adhere to the term bank translations. In Figure 2b and 2d, we see results when we translate the terms only *without the context* in which the term occur.

- Two obvious outliers are cases when translating English to German or Finnish. German and Finnish both have a significant amount of compound words, which has proven to be difficult to translate or rather for language processing in general (Eckman, 1981; Selmer and Lauring, 2015), and the system is expected to translate phrases (in English) to compounds.

- In general, Google's SMT systems outperform NMT when we translate term with context; but NMT performs slightly better in many cases across languages when we translate term without context. This reflects a fundamental difference in the translation mechanisms: in that the NMT end-to-end model pushes the model to translate the sentences holistically, whereas PBSMT systems can handle terms as phrasal units (§3.2).

- Differences in performance among languages are less prominent when we translate without context. This suggests the increased performances among some languages are a result of the language modeling available to the translation system (§3.1.1).

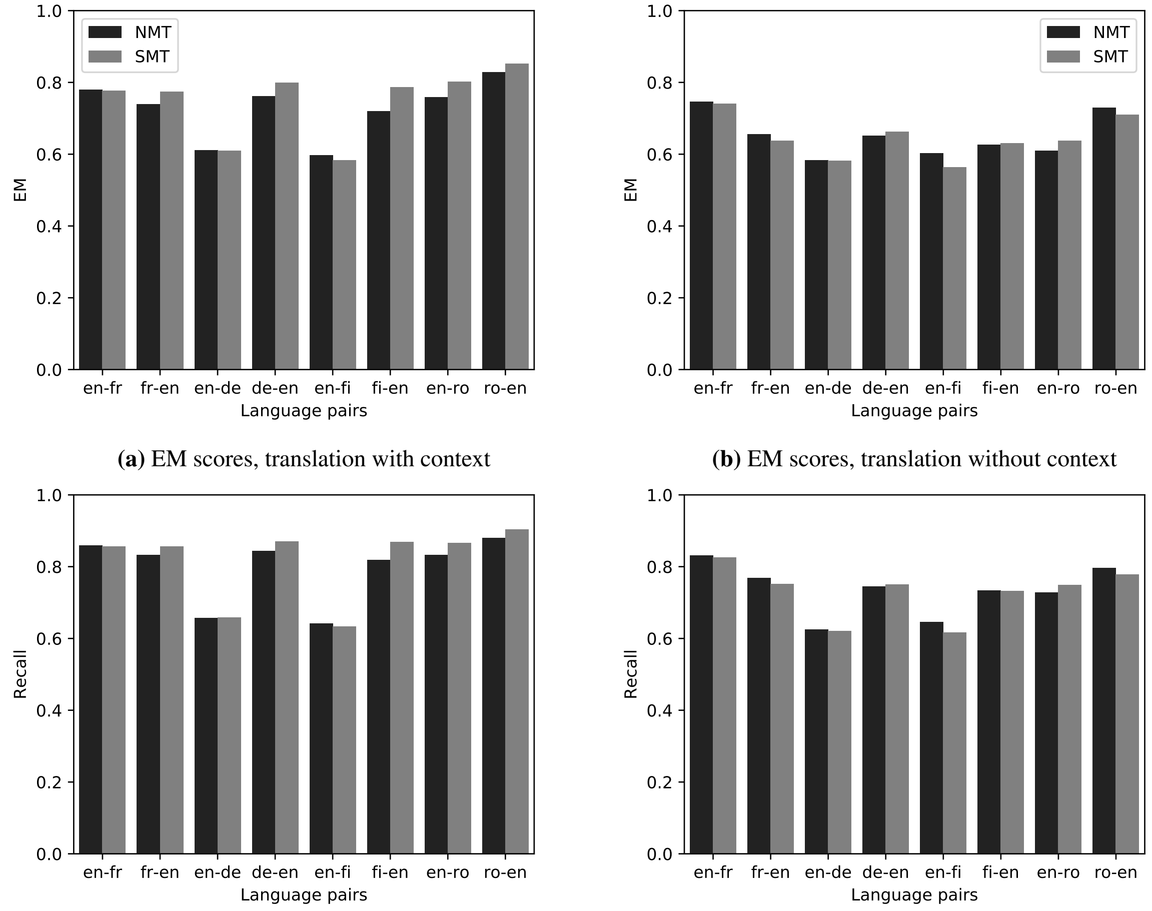
We also conducted a brief analysis of some of the errors we see with regards to term length (number of words) across languages but did not observe significant differences in scores.

5.3 Qualitative Analysis

We do a brief glance at the errors and observe human translation and NMT/SMT among language pairs and directionality.

- For cases where the target language is not English, we observe that NMT are more guilty of paraphrasing not allowed in term translation, like translating "réguler le marché" instead of the correct "réglementer le marché." For English as the target language translation, NMT and SMT both suffer from minor differences that do not affect meaning, suggesting that the English language models are of higher quality.

- English-to-Finnish translation is an outlier in that NMT outperforms SMT when translating *with* or *without* context. We conclude that NMT is better at handling compound words such as "lisäsuojatodistuksen", which is translated from the English multi-word term "supplementary protection certificate."



(a) EM scores, translation with context

(b) EM scores, translation without context

(c) Recall scores, translation without context

(d) Recall scores, translation without context

Figure 2: For the set of sentence pairs where both source and target sentences are validated to contain the source/target term entries as defined in the term bank, we compare scores for NMT v. SMT systems, applying translation either with context or only the source term itself as input.

- For English-Romanian, a lower-resource language pair, we see that NMT is slightly worse-off with or without context. An example of the errors made is translating “self-determination” as “auto-determinarea popoarelor” instead of “autodeterminare a popoarelor”, which is a minor language error unrelated to meaning. We surmise this reflects the zero-shot approach (§3.1.2) in Google’s NMT deemphasizes the nuances of translating resource-poor language pairs.

6 Conclusion

We reach conclusions on NMT/SMT systems compared with human translations that have implications in addition to term validation in translation assurances. Our experiments on translating terms without context show that such MT systems can be useful for the term resource construction process, and can assist translation companies in their work on consolidating terminologies for translators’ referencing purposes.

In the future, we plan to better validate the document-level consistency of terminologies, another key aspect in quality assurance of translation. Specifically, due to the nature of the data applied in this study, we are unable to validate if the terms are *consistently* translated into a single form throughout the whole document. Also, due to our use of publicly-facing commercial MT APIs for our translation models, we have limited insight (based on published work and general knowledge of models) to the inner workings of the systems, and are unable to completely grasp the nature of the training data used by Google in development. This is a trade-off we had to face (as training our own models would be less similar to real-world usage and the model cannot be as extensive due to difficulty in acquiring data), but future work can be based on a balance of both approaches.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 19H05660.

References

Achkasov, Andrei V. 2014. What translators do to terminology: Prescriptions vs. performance. *Journal of Siberian Federal University. Humanities & Social Sciences*, 2:210–221.

Adamska-Salaciak, A. 2010. Examining equivalence. *International Journal of Lexicography*, 23:387–409.

Arango-Keeth, Fanny and Geoffrey S Koby. 2003. Translator training evaluation and the needs of industry quality assessment. In Baer, Brian James and Geoffrey S. Koby, editors, *Beyond the Ivory Tower: Rethinking Translation Pedagogy*, pages 117–134.

Chiocchetti, Elena and Vesna Lusicky. 2017. Quality assurance in multilingual legal terminological databases. *The Journal of Specialised Translation*, 27:164–188.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Eckman, Fred R. 1981. On predicting phonological difficulty in second language acquisition. *Studies in Second Language Acquisition*, 4(1):18–30.

Görög, Attila. 2014. Quantifying and benchmarking quality: the taus dynamic quality framework. *Tradumàtica*, (12):0443–454.

House, Juliane. 2014. Translation quality assessment: Past and present. In *Translation: A multidisciplinary approach*, pages 241–264. Springer.

2010. Iso 29383:2010: Terminology policies – development and implementation. Standard, International Organization for Standardization, Geneva, CH.

Johnson, Ian and Alastair Macphail. 2000. IATE–Inter-Agency Terminology Exchange: Development of a single central terminology database for the institutions and agencies of the european union. In *Proceedings of the Workshop on Terminology resources and computation, LREC 2000 Conference*.

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kageura, Kyo and Elizabeth Marshman. 2019. Translator training evaluation and the needs of industry quality assessment. In O'Hagan, Minako, editor, *The Routledge Handbook of Translation and Technology*, pages 236–331.

Kageura, Kyo. 2012. *The quantitative analysis of the dynamics and structure of terminologies*, volume 15. John Benjamins Publishing.

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

Leitchik, Vladimir M and Serguey D Shelov. 2007. Commensurability of scientific theories and indeterminacy of terminological concepts. *Indeterminacy in terminology and LSP: Studies in honour of Heribert Picht/Ed. by BE Antia. Amsterdam: John Benjamin's Publishing House*, pages 93–106.

Lyding, Verena, Elena Chiocchetti, Gilles Sérasset, and Francis Brunet-Manquat. 2006. The lexalp information system: Term bank and corpus for multilingual legal terminology consolidated. In *Proceedings of the workshop on multilingual language resources and interoperability*, pages 25–31. Association for Computational Linguistics.

Matis, Nancy. 2010. Terminology management during translation projects: Professional testimony. *Lingua-Culture*, 1:107–116.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Peter, Jan-Thorsten, Tamer Alkhouli, Hermann Ney, Matthias Huck, Fabienne Braune, Alexander Fraser, Aleš Tamchyna, Ondřej Bojar, Barry Haddow, Rico Sennrich, et al. 2016. The qt21/himl combined machine translation system. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 344–355.

Sager, Juan C. 1990. *Practical course in terminology processing*. John Benjamins Publishing.

Selmer, Jan and Jakob Lauring. 2015. Host country language ability and expatriate adjustment: The moderating effect of language difficulty. *The International Journal of Human Resource Management*, 26(3):401–420.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Toral, Antonio and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Yamada, Marasu and Nanami Onishi. 2019. Can students still work as a post-editor? In *Proceedings of 25th Annual Meeting of the Japanese Association of Natural Language Processing*, pages 738–741. Japanese Association of Natural Language Processing.