Flagging Comprehensibility Issues in Hindi Text with Question Answering

Anonymous ACL submission

Abstract

There is a critical need for checking the quality of translations while localizing important content across all the industries. This paper presents question-answering based techniques to check the comprehensibility of a text translation. The viability of the method is evaluated using text translated from English to Hindi, where we see comprehensibility issues identified with up to 87% accuracy.

1 Introduction

Those wishing to translate content into local languages use: (i) one or more bi-lingual human translators; (ii) a computer-assisted translation (or CAT) process; or (iii) a completely automated machine translation process. In any of these cases, there are limited options for assessing the quality of the translation. Those hiring bi-lingual translators do not often speak the target language, and machine translation tools do not generally provide any interpretability, risk level, or non-technical assessment information.

Thoroughly checking the quality of a translation thus remains a complex task that requires expert human intervention (Castilho et al., 2018; Stirtz, 2018; Ko, 2011; Ye and Zhang, 2011). As a result, quality checking slows the pace of localizing important text content (e.g., health, religious, legislative, or legal information), because the quality of such translations needs to be verified before publication (Ramos, 2020; Kmiecicka, 2021; Ghobadi et al., 2017). COVID-19 has uniquely exposed these translation checking inefficiencies. The demand for translation of healthcare, medical research and pharmaceutical content is increasing, but the there are limited expert resources available (El-Jardali et al., 2020; Anastasopoulos et al., 2020; Way et al., 2020; Dhawan et al., 2021).

This paper presents a method to augment translation checking by flagging potentially poor quality translations as related to "comprehensibility." The method probes both the informational content of a translation and the inferences related to the intended meaning. It does so by automatically answering questions paired with both the source content and an automatic back translation of the target, translated content. Answers to these questions reveal the level of difficulty for an uninitiated native speaker to arrive at the intended meaning.

We demonstrate the method using "draft" Hindi [hin] translations of English content [eng]. Certain of these translations are manually perturbed to create known translation issues, while others have been quality checked by humans. We find that it is feasible to evaluate the comprehensibility of the draft Hindi text using both: (i) pre-trained Hindi question answering models operating on the draft Hindi text; and (ii) pre-trained English question answering models operating on automatic back translations of the draft Hindi text.

2 Related Work

Question and answer is a common format used to test students in reading comprehension. There are a variety of these kinds of tests that measure a variety of different comprehension skills (Keenan et al., 2008). The QRI reading comprehension test (Leslie and Caldwell, 1990), for example, involves reading stories aloud and then having students respond to short-answer comprehension questions. Similar tests may involve a combination of multiple-choice questions about a passage or a retelling exercise. Of course, all of these kinds of tests are performed manually and are typically employed in an educational context.

Automatic or machine question answering (sometimes referred to as simply "question answering", "machine comprehension", or "reading comprehension") is the task of answering questions given one or more text passages as context. The performance of neural network based models trained on this task has increased considerably in recent

years. This progress is in large part due to the release of large language models like BERT (Devlin et al., 2019) and new data sets that have introduced "impossible" questions (Rajpurkar et al., 2018), bigger scales (Kwiatkowski et al., 2019), and various forms of context. As described in a recent paper by Sen and Saffari (2020), question answering models have outperformed human baselines on the widely-used SQuAD 1.1 and SQuAD 2.0 data sets (Rajpurkar et al., 2016, 2018).

Despite this clear progress in automated question answering, the authors are not aware of any attempt to apply question answering models to automatically assess the quality of translations.

3 Methodology

See Figure 1 for an overview of our question answering based approach to checking the "comprehensibility" of a draft translation. In our approach, content in the source language is paired with one or more curated comprehension questions that should be directly answerable from the respective passage. After the source language content is translated into the target language, a machine translation model is used to automatically back translate the target language content back into the source language. A question answering model then answers the questions using both the original and back translated content. Answers corresponding to the original content are compared with answers from the back translation to flag possible comprehensibility issues.

To compare the answers from the back translation with the answers from the original content, we utilize a pre-trained text embeddings. Each answer is vectorized using the embeddings and a similarity between the embeddings is calculated using cosine similarity. A simple threshold is used to flag answers that are dissimilar to that generated from the original content.

In certain cases, supplemental source language content is utilized to increase flagging performance. That is, answers to the curated comprehension questions are generated from the source content along with a variety of other source language references, where the other source language references contain similar information. The answer from the back translated candidate is then compared with answers from the original source content and the various other source language references. This addition of supplemental references helps to ensure that synonyms in the back translation do not impact the flagging as significantly.

When both source content and supplemental source language references were used, we experimented with two slightly different flagging methods based on calculated similarities between the various answers. In a first method, we take the average of similarity scores for all candidate answer to source language content answer pairs. Alternatively, we flag the candidate answer if any one similarity score is below a threshold. The accuracy of each flagging method was calculated using the data discussed in Section 4. In the end, the results reported here use the average-based method, which generally performed best.

4 Experiments

In our experiments, we compare the utility of flagging poor quality translations using automatic question answering: (i) in the source language using a back translation of the draft or candidate forward translation; and (ii) natively in the target language using the draft translation itself along with a reference (or "model") translation(s) in the target language. Comparing these two scenarios will allow us to understand how the performance of the method degrades due to noise in the back translation.

To this end, we assembled a dataset[1] of English and Hindi Bible passages. Hindi in treated as the target language and English is treated as the source language. Specifically, this dataset includes:

- Multiple versions of 300 English [eng] Bible passages (NET, CSB, NIV, NLT, GNT, ESV, KJV, WEB, and NAS)
- 300 comprehension questions in English paired with those Bible passages
- Multiple versions of the corresponding Hindi [hin] Bible passages (HIN2017 from bible.com, the Hindi Holy Bible from wordproject.org, and the Sab Ki from bible.is)
- The corresponding comprehension questions in Hindi paired with those Bible passages

The comprehension questions were sourced from SIL International's Transcelerator project[2], which includes real world question and answer

[1] to be released publicly on publication
[2] https://github.com/sillsdev/Transcelerator

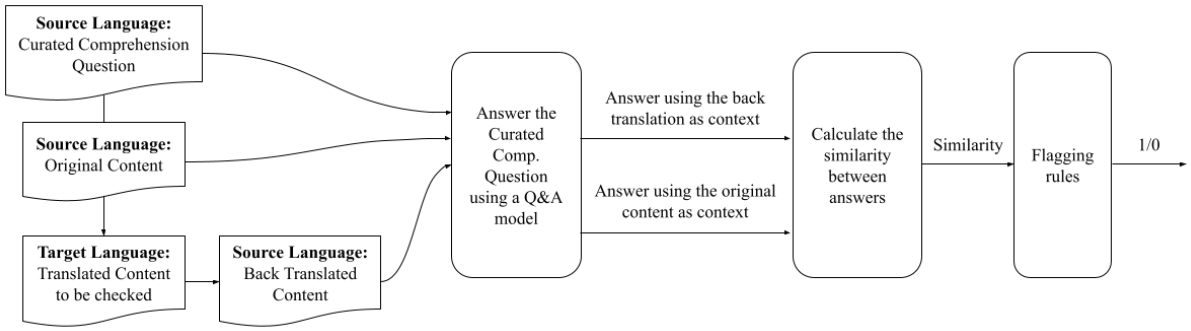

Figure 1: The generally methodology for checking the "comprehensibility" of a translation using question answering.

pairs used to check the comprehensibility of draft Bible translations.

To evaluate the comprehension checking method of Section 3 natively in the target language, the HIN2017 translation is treated as the "draft" translation. We manually inject comprehensibility issues in 25% of this HIN2017 data to create known comprehensibility issues. The injected issues include under-specifying context (e.g., removing an important clause), changing words, and jumbling words, all of which are likely real world scenarios in a translation process. The other 75% of the HIN2017 data is considered to be good quality given that it was already quality checked and published by the Bible publisher. The Sab Ki and Hindi Holy Bible data is left unperturbed to be used as supplemental reference content (see Figure 2).

To evaluate the comprehension checking method of Section 3 in the source language using a back translation, we back translate the HIN2017 data into English using Google Translate[3]. In English, one of the Bible translations is treated as the source content and the remaining are treated as reference content (see Figure 2).

We use the bert-multi-cased-finetuned-xquadv1[4] from Hugging Face to answer questions in Hindi, and we use the DistilBert[5] model (fine-tuned on SQuAD data) to answers questions in English. To calculate similarity between answers we use pre-trained LaBSE (Feng et al., 2020) and LASER[6] text embeddings. As a further baseline, we use BLEU scores (Papineni et al., 2002) to calculate a raw n-gram based similarity between answers. A threshold of 0.675 was used for all experiments, where similarities below this threshold were considered to be anomalous.

[3]https://translate.google.com/

[4]https://huggingface.co/mrm8488/bert-multi-cased-finetuned-xquadv1

[5]https://huggingface.co/distilbert-base-uncased-distilled-squad

[6]https://github.com/facebookresearch/LASER

5 Results

Table 1 shows the results of question answering based flagging of Hindi comprehension issues. Using the method of Section 3 we see accuracy metrics higher than 80% when flagging comprehension issues with either a Hindi question answering model or an English question answering model. In fact, the best performing scenario that we evaluated used automatic back translations and question answering in English. This demonstrates that the higher performance of the English question answering model outweighs any information loss from using a machine translation model to back translate the draft. Such a results suggests that the method could be viable for a wide range of languages, because a question answering model in the target language is not required as long as you can back translate the target content.

As mentioned in Section 3, we used two methods to flag comprehension questions given multiple reference contexts. (OR) in Table 1 represents flagging when any answer pairs are below the similarity threshold and (AVG) represents flagging only when the average score is below the threshold. Generally, the average-based flagging performed best, which suggests that having a diversity of supplemental references can boost performance.

6 Conclusion & Future Work

Using question answering, we demonstrate that comprehensibility-related issues in Hindi draft translations can be identified automatically with an accuracy up to 87%. This method seems to be viable for other languages as well, because a question answering model in the target language

Language	Flagging	Precision	Recall	Acurracy
hin	BLEU (AND)	0.42	0.58	0.70
hin	BLEU (AVG)	0.42	0.78	0.68
hin	LaBSE (OR)	0.49	0.80	0.75
hin	**LaBSE (AVG)**	**0.72**	**0.77**	**0.87**
eng	LaBSE (OR)	0.38	0.95	0.61
eng	**LaBSE (AVG)**	**0.59**	**0.84**	**0.82**

Table 1: Results of questions answering based flagging of Hindi comprehension issues. The BLEU score and LaBSE embedding based similarity flagging methods are shown here, because they represent a naive baseline approach and the best performing approach. (OR) in the table indicating flagging if any of the similarity scores are below the threshold and (AVG) indicates flagging if the avergae similarity score is below the threshold.

is not required as long as a back translation can be produced. In the future, we would like to further validate the methodology using data from other domains, such as healthcare. We would also like to test the method with a variety of other question answering models and with data from lower resourced languages.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzm'an, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Meng Niu, Alp Oktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. Tico-19: the translation initiative for covid-19. *ArXiv*, abs/2007.01788.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and J. Moorkens. 2018. Approaches to human and machine translation quality assessment.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Natasha Dhawan, Ishwaria M. Subbiah, Jonathan C Yeh, Benjamin Thompson, Zachary Hildner, Areeba Jawed, Eric Prommer, and Christian T Sinclair. 2021. Healthcare disparities and the covid-19 pandemic: Analysis of primary language and translations of visitor policies at nci-designated comprehensive cancer centers. *Journal of Pain and Symptom Management*, 61:e13 – e16.

Fadi El-Jardali, Lama Bou-Karroum, and Racha Fadlallah. 2020. Amplifying the role of knowledge translation platforms in the covid-19 pandemic response. *Health Research Policy and Systems*, 18.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Mehdi Ghobadi, Golnaz Madadi, and Bahareh Najafian. 2017. A study of the effects of time pressure on translation quantity and quality. *International Journal of Comparative Literature and Translation Studies*, 5:7–13.

Janice M. Keenan, Rebecca S. Betjemann, and Richard K. Olson. 2008. Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12:281 – 300.

Eliza Kmiecicka. 2021. Mistakes in specialist translations and their possible consequences in the legal communication.

Leong Ko. 2011. Translation checking: a view from the translation market. *Perspectives*, 19:123 – 134.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Lauren Leslie and JoAnne Schudt Caldwell. 1990. *Qualitative reading inventory*. Harper Collins New York.

Kishore Papineni, 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*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for

	Question	Contexts
eng	What did the angel tell Joseph to name the child?	She will give birth to a son, and you will name him Jesus, because he will save his people from their sins.
		She will give birth to a son, and you are to give him the name Jesus, because he will save his people from their sins.
		...
		She will bear a son, and you shall call his name Jesus, for he will save his people from their sins.
hin	उसे पुत्र का क्या नाम रखना चाहिए?	वह पुत्र जनेगी अपने लोगों का उनके पापों से उद्धार करेगा।
		वह पुत्र जनेगी और तू उसका नाम यीशु रखना; क्योंकि वह अपने लोगों का उन के पापों से उद्धार करेगा।
		वह एक पुत्र को जन्म देंगी. तुम उनका नाम येशु रखना क्योंकि वह अपने लोगों को उनके पापों से उद्धार देंगे.

Source Reference/Supplement Draft/Candidate

Figure 2: Example data from our experiments for a single Bible passage, which includes multiple contexts in Hindi, multiple contexts in English, a question in Hindi, and a question in English.

machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Fernando Prieto Ramos. 2020. Facing translation errors at international organizations: What corrigenda reveal about correction processes and their implications for translation quality. *Comparative Legilinguistics*, 41:133 – 97.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *EMNLP*.

Timothy M Stirtz. 2018. Linguistic checks: The missing link in modern language development. *SIL Electronic Working Papers*, 127.

Andy Way, Rejwanul Haque, Guodong Xie, Federico Gaspari, Maja Popovic, and Alberto Poncelas. 2020. Rapid development of competitive translation engines for access to multilingual covid-19 information. *Informatics*, 7:19.

Na Ye and Guiping Zhang. 2011. Study on the impact factors of the translators' post-editing efficiency in a collaborative translation environment. In *MTSUMMIT*.