Quality Estimation of Machine Translated Texts based on Direct Evidence Approach

Vibhuti Kumari and **Kavi Narayana Murthy**
School of Computer and Information Sciences, University of Hyderabad
vibhuroy9711@gmail.com,knmuh@yahoo.com

Abstract

Quality Estimation task deals with the estimation of quality of translations produced by a Machine Translation system without depending on Reference Translations. A number of approaches have been suggested over the years. In this paper we show that the parallel corpus used as training data for training an MT system holds direct clues for estimating the quality of translations produced by that MT system. Our experiments show that this simple, direct and computationally efficient method holds promise for quality estimation of translations produced by any purely data driven machine translation system.

1 Introduction

The performance of Machine Translation (MT) systems is measured either using Manual Evaluation, using metrics such as Adequacy and Comprehensibility, or using automatic methods, using metrics such as BLEU and TER, by comparing with Reference Translations. Maurya et al. (2020) Quality Estimation (QE), on the other hand, deals with automatic estimation of quality of translations produced by an MT system without using Reference Translations. Machine translation generally works sentence by sentence and the primary goal of the Quality Estimation task is also to measure the quality of translations at sentence level.

QE of MT outputs has several benefits. Good translations can be selected, post-edited as required and added to the training data. Poor quality instances can be removed from training data to reduce noise. QE helps in more accurate estimation of post-editing time and effort and in taking associated decisions in commercial translation.

Parallel Corpus Data required for building MT systems is generated today mostly using automatic methods such as web crawling and back translation, and the data so generated is usually quite noisy. Modern Deep Neural Architectures are data hungry. Quantity and quality of data are thus both very important. Neural MT (NMT) performance, for example, has been shown to be highly dependent on the size of the training data (Koehn and knowles, 2017) as well as the quality (Khayrallah and Koehn, 2018; Batheja and Bhattacharyya, 2022).

A large number of techniques have been proposed for quality estimation. The annual Workshop on Machine Translation (WMT) has been including a shared task on quality estimation for many years now. A typical approach is to use manually scored translations as training data for training a Machine Learning system to perform the QE task.

More recently, Sentence Embeddings generated using deep learning neural network architectures such as Transformer, have been used to estimate the quality of translations. The core idea is to compare the sentence embeddings of the source language sentence with that of the target language sentence. Higher the similarity of sentence embeddings, higher is the semantic similarity of the sentences. Language agnostic BERT Sentence Embedding model (LaBSE) (Feng et al., 2022) is an example of this approach.

In this paper, we propose a much simpler and more direct approach called the Direct Evidence approach. We show that while being computationally much less expensive, our method correlates well with other more sophisticated methods such as LaBSE.

While the research in MTQE is rich in terms of ideas, techniques, tools and resources, it appears that none of them are directly looking at the parallel corpus that is used for building MT systems, for clues about quality of translations. Here we propose what we call Direct Evidence approach, which is based directly and solely on the training data that is used to build MT systems. No other training data or resource is required. This approach is also conceptually simple and computationally very efficient.

2 Related Work

Till recently, the mainstream approach to QE was to use a set of manually scored translations to train a machine learning system. Most of the works reported in the WMT shared tasks on QE used this approach. Several sub-tasks and related tasks were also taken up in the WMT workshops. Word level QE deals with marking of words as OK or BAD. In fact, sentence level scores were often computed or estimated using these word level scores. Scoring entire documents was another task. Identifying Source Language (SL) words that cause quality issues was also looked at. Explainable QE task and Critical Error Detection task were included in the WMT-2022 conference. Both Direct Assessment on post-edit data (called MLQE-PE) and Multidimensional Quality Metrics (MQM) were included. In the prevalent evaluation practices, QE systems were assessed mainly in terms of their correlation with human judgements. Zerva et al. (2022) describe the findings of the 11th edition of the QE shared task held as part of WMT-2022. Participants from 11 different teams submitted altogether 991 systems to different task variants and language pairs in WMT-2022. Zaretskaya et al. (2020) ask whether the current QE systems are useful for MT model selection. Gladkoff et al. (2022) focus on the amount of data that is required to reliably estimate the quality of MT outputs. They use Bernoulli Statistical Distribution Modeling and Monte Carlo Sampling Analysis towards this end. Don-Yehiya et al. (2022) focus on quality estimation of machine translation outputs in advance. They present a new task named PreQuEL, the task of predicting the quality of the output of MT systems based on the source sentence only. Blain et al. (2023) summarize the findings of the 2023 edition of the WMT conference. Khayrallah and Koehn (2018) explore how various types of noise in the training data impact the quality of neural machine translation systems. They find that neural models are generally affected more by noise than statistical models. NMT performance degraded by 9.9 BLEU points when noise was added while Statistical Machine Translation (SMT) actually gained 1.2 BLEU points. Effect of various types of noise such as misaligned sentences, misordered words, wrong language, untranslated or very short segments and raw crawl data have been explored. Feng et al. (2022) introduce Language agnostic BERT Sentence Embedding (LaBSE), and compare with LASER (Language Agnostic Sentence Embedding Representations) Artetxe and Schwenk (2019) and m-USE (Multilingual Unsupervised and Supervised Embeddings) Yang et al. (2019) approaches. LaBSE outperformed LASER and m-USE in many scenarios. LaBSE is a multilingual sentence embedding model for more than 109 languages based on dual encoder transformer architecture of BERT Devlin et al. (2018); Vaswani et al. (2017). All the above three models are computationally highly expensive. For example, LaBSE uses the BERT Base encoder architecture with 12 transformer blocks, 12 attention heads, 768 per-attention hidden units. Sentence embeddings are extracted as L2 Normalized [CLS] token representation from the last transformer block. Models are trained on Cloud TPU V3 with 32-cores using a global batch size of 4096, with a maximum sequence length of 128, using AdamW optimizer with initial learning rate e^{-3} and linear weight decay. The default margin value for additive margin softmax is set to 0.3. Batheja and Bhattacharya (2023) introduce a few-shot transfer learning based approach to QE and show that using this approach for corpus filtering gives higher improvements in MT performance compared to LaBSE based corpus filtering. Bane et al. (2022) explore various data filtering methods and evaluate them on the downstream task of NMT. They conclude that cross entropy based filtering outperforms other approaches. Taghipour et al. (2011) view corpus refinement as an Outlier Detection task. In order to detect and remove the mistranslations in a parallel corpus, they map each sentence pair into an N-dimensional feature space and then estimate the density for each one of them. The least dense points are treated as outliers and are removed from the corpus. Gala et al. (2023) present IndicTrans2, claiming to provide high quality and accessible MT models for all the 22 scheduled Indian languages. They release BPCC, a parallel corpus including a total of 230 M bitext pairs, of which about 126 M were newly added in this release, including 644 K manually translated sentence pairs. They also released the first n-way parallel benchmark covering all 22 Indian languages. Das et al. (2024) aim to remove the incorrect translations from the dataset to make the translation quality better. Sentences with poor translation quality (BLEU score lesser than a threshold) are treated as noise and discarded from the dataset. Xu et al. (2019) propose a novel approach to filter this noise from synthetic data. For each sentence pair of the synthetic data, they compute a semantic similarity score using bilingual word embeddings. Xu and Koehn (2017) propose a fast and scalable data cleaning system for noisy web-crawled parallel corpora. They propose a novel type of bag-of-words translation features, and train logistic regression models to classify good data and synthetic noisy data in the feature space. Aulamo et al. (2020) introduce OpusFilter, a flexible and modular toolbox for filtering parallel corpora. In contrast to tools such as bicleaner Sánchez-Cartagena et al. (2018) and Zipporah Xu and Koehn (2017) that implement a single method for parallel corpus filtering, OpusFilter is designed as a toolbox that is useful for testing and using many different approaches. Cui et al. (2013) propose a graph based random walk approach to clean bilingual data for SMT. A PageRank-style random walk algorithm Brin and Page (1998); Mihalcea and Tarau (2004); Wan et al. (2007) is used to iteratively compute the importance score of each sentence pair that indicates its quality: the higher the better. Unlike other data filtering methods, their proposed method utilizes the importance scores of sentence pairs as fractional counts to calculate the phrase translation probabilities based on Maximum Likelihood Estimation. Sloto et al. (2023) present the findings of the WMT 2023 shared task on Parallel Data Curation. They pose the open-ended shared task of finding the best subset of possible training data from a collection of Estonian-Lithuanian web data. Junczys-Dowmunt (2018) introduce dual conditional cross-entropy filtering for noisy parallel data. For each sentence pair of the noisy parallel corpus they compute cross-entropy scores according to two inverse translation models trained on clean data. They penalize divergent cross-entropies and weigh the penalty by the cross-entropy average of both models. Sorting or thresholding according to these scores results in better subsets of parallel data. Wu et al. (2024) explore how prompt strategies affect downstream translation performance. Then, they conduct extensive experiments with two fine-tuning methods, three LLM backbones and 18 translation tasks across nine language pairs. Their findings indicate that in some cases, these specialized models even surpass GPT-4 in translation performance, while they still significantly suffer often from off-target translation issue, even if they are exclusively fine-tuned on bilingual parallel documents.

3 Direct Evidence Approach

Translation is a meaning preserving transformation of texts from a Source Language (SL) to a Target Language (TL). This is generally done sentence by sentence, or more generally, segment by segment. In order to preserve the meaning of the SL sentence, words and phrases in SL sentences need to be mapped to equivalent words and phrases in the TL. Other aspects of syntax and semantics such as agreement, word order, semantic compatibility will also need to be addressed. Modern purely data driven approaches such as Statistical Machine Translation and Neural Machine Translation are based on the view that all linguistic regularities and idiosyncrasies are indirectly present in the parallel corpus and parallel corpus alone is sufficient, no other data or linguistic resource is needed. A Machine Translation system can be obtained by training on a training data set consisting of a parallel corpus alone.

We believe that the training data also has clues useful for estimating the quality of translations produced by the MT system. In particular, here we focus on lexical transfer. We show that the Word Co-occurrence Matrix (WCM) holds direct clues for estimating the quality of lexical transfer and hence quality of translation as a whole.

Statistical basis for performing lexical transfer comes mainly from word co-occurrence statistics. Let SL-TL be a parallel corpus consisting of n Source Language segments $S_1, S_2, S_3, ..., S_n$, paired with their translational equivalents $T_1, T_2, T_3, ..., T_n$ in the Target Language. We say SL word i co-occurs with TL word j if the TL word j occurs anywhere in the translational equivalent of a SL sentence in which the word i occurs. Let V_s be the Vocabulary of the Source Language (total number of distinct word forms occurring in any of the SL segments) and V_t be the Vocabulary of the Target Language. Then Word Co-Occurrence Matrix WCM is a V_s x V_t matrix of non-negative integers where $WCM_{i,j}$ indicates the total number of times the Source Language word i had co-occurred with the Target Language word j in the entire training data set. Clearly, WCM will be a very large and very sparse matrix.

A large $WCM_{i,j}$ value indicates a strong correlation between the SL word i and TL word j in the training corpus. If an SL word i co-occurs with a TL word j large number of times, if i does not occur with too many other TL words with high frequency, if the WCM counts for other possible mappings in TL are significantly lower, all these indicate that the lexical transfer of i to j during translation can be done with high confidence. When the evidence in the form of co-occurrence counts coming from the training data is weak, the MT system may still go ahead and substitute the word j for word i based on the combined evidence coming from other parts of the sentence, language model, etc. This may be an optimal decision taken by the MT system with regard to some specified loss function. Optimal choice in some probabilistic sense may not be the correct choice, it may just be the best of several possible choices, none of which may be correct. MT systems generally go ahead and produce translations, whether they are sure or not-so-sure or not-at-all-sure.

Here we hypothesize that *the fraction of words in a SL sentence that have strong co-occurrence relations with any of the words* *in the TL sentence produced by the MT system, is a direct indicator of quality of translation. We call this Direct Evidence Approach.*

4 Methodology

First we compute the Word Co-occurrence Matrix WCM from the MT training data. Then for each SL-TL sentence pair in the test set, we check the number of SL words (excluding stop words) for which there is 'strong evidence' in the training data. A source language word W is said to have 'strong evidence' if it co-occurs at least T times with any of the target language words in the sentence pair, where T is a specified Threshold. We call the percentage of SL words with 'strong evidence' as the DE Score for the sentence pair. DE Scores range from 0 to 100. DE Score is taken as a measure of quality of translation.

5 Experiments and Results

Experiment 1

In our first experiment we use an English-Kannada parallel corpus consisting of 4,014,931 segments (that is approximately 4 Million segments) (Ramesh et al., 2021) There are about 36M tokens in English and 27M tokens in Kannada. The Vocabulary size for English is 281,881. Only 42,222 (less than 15%) occur at least 20 times. 78.5% of words occur less than 10 times, 69% of words occur less than 5 times, 44.47% of words occur only once. This highly skewed distribution of words in all human languages is very well understood and expressed through laws such as Zipf's law (Zipf, 1949) and Mandelbrot's law (Mandelbrot, 1965). The Vocabulary of the Kannada part is 1,253,589. This number is larger due to the much more complex morphology we see in Dravidian Languages such as Kannada. Only 82,227 (6.5%) occur at least 20 times. 89.2% of words occur less than 10 times, 81.8% of words occur less than 5 times, 57.8% of words occur only once. The general picture will be similar for any pair of languages in the world.

If a SL word i occurs only once and the translation of the sentence in which it occurs has n words, then i can be mapped to any one of these n TL words with equal probability. While an MT system may use other clues such as mappings of other words in the SL sentence and language model probabilities, it will still be decision that is not based on very strong evidence. Low frequency words show poor co-occurrence relations and hence less statistical evidence for lexical transfer. Low frequency words are large in number in any language and this is a big issue for any purely data driven model. Larger data is better but whatever may be the size of the data, the problem remains pertinent.

Very high frequency words can also pose challenges. They usually include determiners, prepositions and other function words. Words such as 'the', 'of', 'by' occur with very high frequency in English, none of them map directly to any word in Kannada. WCM will show large number of possible mappings, all (or almost all) of them will be wrong. This is again a hopeless situation. Phrase based approaches and sub-word models attempt to address these problems and are successful to some extent.

Keeping these ideas in mind, we build WCM for words that co-occur at least 20 times in the training set, we exclude words which occur more than 10,000 times in the corpus. Under these assumptions, WCM matrix can be built very fast (it took less than 4 minutes on a 40 core Intel Xeon Silver 4114 CPU at 800 MHz server) and the size of uncompressed the WCM file is only 44 MB. There are 1,474,792 entries in the WCM matrix, there are only 38,502 English words in this matrix.

We divide the corpus into training, development and test sets with 4,004,894, 5000 and 5037 segments respectively and train an SMT system using MOSES (Koehn

DE Score	No. of Segments	BLEU Score
< 20	847	6.33
< 30	2082	6.78
< 40	3036	7.06
< 50	3588	7.46
≥ 50	1449	9.16
≥ 60	669	10.49
≥ 70	327	11.34
≥ 80	237	10.80

Table 1: DE Scores vs. BLEU Scores for English-Kannada

et al., 2007). WCM is computed for the training set.

We run the trained SMT system on test data. We compute the DE Scores as described above for each segment, taking the Threshold T as 20. We pick out SL-TL pairs from the test data as also from the generated MT outputs based on selected ranges of DE Scores. Taking the TL part in the test data as Reference, we compute BLEU scores: See Table 1.

We can clearly see a positive correlation between the DE Scores we obtained and the BLEU scores, up to a threshold of 70. Manual observations also clearly showed the gradation in quality of translations correlating with the DE Scores we compute. Sentences which got high DE Scores were generally of much better translation quality compared to sentences which got a poor DE Score.

Experiment 2

Next we compute sentence level BLEU scores and look for correlation between these BLEU scores and the DE Scores. Over 5037 segments of test data, we get a Pearson Correlation Coefficient of 0.209405. The p-value is < 0.00001 Hence the result is significant at the typical $p < 0.05$.

Experiment 3

Training corpora used for building MT systems are often not available for us to experiment with. Here we take up one case where we could locate the training data as also the MT outputs and Reference Translations. This relates to English-Hindi SMT system developed by Piyush Dungarwal et al from IIT Bombay (Dungarwal et al., 2014) in the Ninth Workshop on SMT, WMT-2014. Training data consists of 273,885 segment pairs, including 3,378,341 tokens in English and 3,659,840 tokens in Hindi. There are 129,909 unique word forms in English, of which only 19,100 occur 10 times or more. Total number of unique word forms in Hindi is 137,089, of which only 18,587 occur 10 times or more. In English, 30 words occur with frequency more than 10,000 and are taken as frequent words in our experiments. In Hindi, there are 33 very high frequency words. These high frequency words are excluded from WCM computations. This makes the WCM matrix smaller and saves time too. The WCM matrix could be computed in a minute or so on an ordinary Desktop computer. The WCM has 642,341 entries including 242,477 pairs that co-occur ≥ 20 times.

There are 2507 segments in each of the test set source, MT system output, and Reference Translations. We compute the DE Scores based purely on the WCM matrix which is based only on the training corpus. We extract subsets of the MT outputs and corresponding reference translations based on the DE Score ranges. The BLEU scores are as shown in Table 2.

Here again we see a clear gradation in BLEU scores correlating with the DE Scores. Higher the DE Score, higher the BLEU.

The results of these preliminary experiments support our claim that the clues needed for MT QE are present in the training data itself, nothing else may be necessary. We do not even need an MT

DE Score	No. of Segments	BLEU Score
< 50	133	6.00
≥ 50	2374	10.36
≥ 60	2154	10.61
≥ 70	1733	10.93
≥ 80	1120	11.69
≥ 90	463	12.47

Table 2: DE Scores and BLEU Scores for English-Hindi

system to predict the quality of translations it will produce, just the training data is sufficient.

Experiment 4

We then calculated the DE-Scores for the 4 Million segment Training Data used for building our English-Kannada SMT system as described in Experiment 1. Figure 1 shows the histogram plot of DE-Scores obtained. It can be observed that a significant part of the training data got DE-Scores less than 50, many cases even less than 10. This can be useful in locating and reducing noise in the training data.

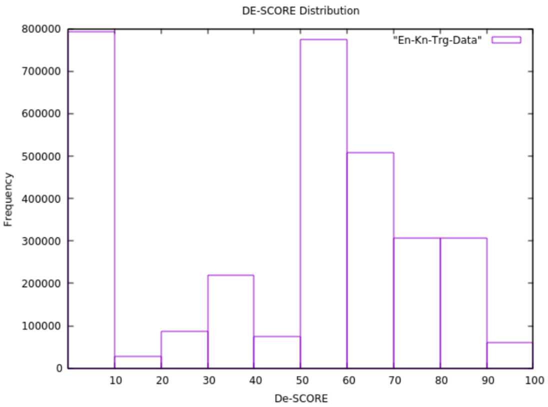

Figure 1: DE-Scores for English-Kannada Training Data Set

Experiment 5

LaBSE score $\geq$	DE $= 100$	DE ≥ 90	DE ≥ 80	DE ≥ 70
0.80	85.35%	90.77%	96.33%	97.77%
0.85	85.51%	91.46%	96.74%	98.05%
0.90	85.14%	92.20%	96.95%	98.12%
0.95	85.62%	90.88%	94.51%	95.83%

Table 3: DE-Score ranges for high LaBSE scores

DE Score $\geq$	LaBSE ≥ 0.80	LaBSE ≥ 0.85
70	91.42%	71.04%
80	91.53%	71.22%
90	91.65%	71.56%
100	91.37%	70.94%

Table 4: LaBSE scores for high DE-Scores

In our final experiment, we compare DE-Scores with LaBSE scores on a large back translated English-Hindi corpus (Gala et al., 2023). This corpus has about 37 Million sentence pairs. We compute the DE-Scores from a Word Co-Occurrence Matrix built over the entire corpus, taking out only a small number of stop words (273 words for English and 338 Words for Hindi). The WCM file has about 14M entries. DE-Scores are rational numbers (ratio of integers) and thus discrete. The range is 0 to 100. LaBSE scores, on the other hand, are floating point numbers in the range -1 to +1. So we avoid directly computing Correlation Coefficients.

First we check what percentage of sentences get a DE-Score of 70 or more, when the LaBSE score is, say, above 0.8. We find that 97.77% of the sentences which had a LaBSE score of 0.8 also have a DE-Score of over 70. The table below shows results for other high end LaBSE scores and corresponding DE-Scores. It can be seen that whenever the LaBSE score if high, DE-Score also tends to be high.

Reversing the question, we also check the LaBSE scores when DE-Scores are high. The following table again shows that whenever DE-Scores are high, LaBSE scores also tend to be high.

Computing DE-Scores is computationally much less expensive compared to

LaBSE and other scores based on deep learning based sentence embeddings. We can use DE-Scores at least to select better quality translations.

We also observe that only 258 sentences in the corpus have a negative LaBSE score. 93.7% of these cases, DE-Score is also zero.

6 Conclusions

In this paper we hypothesize that the Parallel Corpus used for Training an MT system holds clues about the quality of translations the MT system can produce. We propose a simple and direct approach to quality estimation based solely on the training data. A word co-occurrence matrix is constructed from the training corpus and used to estimate the sentence by sentence quality of translations. Each sentence gets a score called DE Score, which is indicative of the quality of translation. Manual observations show that good quality translations generally tend to get higher DE Scores and poor quality translations tend to get lower scores. Our experiments reconfirm this. This simple and direct evidence approach to MT Quality Estimation appears to holds promise. We can estimate the quality of translations even without / before running the MT system. We do not need any other data or resource, we only need the training corpus.

DE-Scores provide us a spectrum of quality grades and since they are based on co-occurrence counts, Out of Vocabulary (OOV) words are only cases that lie just outside the low end of this spectrum.

Missing words automatically get reflected in poor DE Scores but extra words in TL can be detected by performing a TL to SL WCM check. If large scale manual post-edit data such as HTER scores are available, then we can estimate the various thresholds using machine learning techniques instead of using human judgement as we have done here.

In summary, DE Score is a simple, direct and efficient method for the QE task. We can also easily deal with untranslated text, third language, emojis and other invalid characters etc.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Mikko Aulamo, Sami Virpioja, and Jorg Tiedemann. 2020. Opusfilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 150–156.

Fred Bane, Celia Soler Uguet, Wiktor Stribizew, and Anna Zaretskaya. 2022. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Bienniel Conference of the Associations for Machine Translation in the Americas Oriando, USA*, pages 313–325.

Akshay Batheja and Pushpak Bhattacharya. 2023. "a little is enough":few-shot quality estimation based corpus filtering improves machine translation. In *Findings of the Association for Computational Linguistics, ACL 2023*, pages 14175–14185.

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving machine translation with phrase pair injection and corpus filtering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5395–5400.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, Jose G.C. de Souza, Beatriz Silva, Tania Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and Andre F. T.Martins. 2023. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation(WMT)*, pages 629–653.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30:107–117.

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for smt using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 340–345.

Sudhansu Bala Das, Leo Raphael Rodrigues, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2024. An approach for mistranslation removal from popular dataset for indic mt task. arXiv:2401.06398.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. Prequel: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183.

Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2014. The IIT Bombay Hindi-English translation system at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 90–96, Baltimore, Maryland, USA. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring uncertainty in translation quality evaluation (TQE). In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1454–1461, Marseille.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. *arXiv preprint arXiv:1809.00197*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. 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*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Phillipp Koehn and Rebecca knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Benoit Mandelbrot. 1965. Information theory and psycholinguistics. In B B Wolman and E Nagel, editors, *Scientific psychology*. Basic Books.

Kaushal Kumar Maurya, Renjith P Ravindran, CH Ram Anirudh, and Kavi Narayana Murthy. 2020. Machine translation evaluation: Manual vs. automatic - a comparative study. In *Data Engineering and Communication Technology*, volume 1079 of *Advances in Intelligent Systems and Computing*, pages 541–553. Springer.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. arXiv cs.CL 2104.05596.

Víctor M Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the third conference on machine translation: shared task papers*, pages 955–962.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the wmt 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation(WMT)*, pages 95–102.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *In MT Summit XIII. Machine Translation Summit(MT Summit-11)*, pages 19–23.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 552–559.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Guanghao Xu, Youngjoong Ko, and Jungyun Seo. 2019. Improving neural machine translation by filtering synthetic parallel data. *Entropy*.

Hainan Xu and Phillip Koehn. 2017. Zipporah:a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Anna Zaretskaya, José Conceição, and Frederick Bane. 2020. Estimation vs metrics: is QE useful for MT model selection? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 339–346, Lisboa, Portugal. European Association for Machine Translation.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

G K Zipf. 1949. *Human Behaviour and the Principle of Least effort: An introduction to Human Ecology*. Addison-Wesley.