Quality Scoring of Source Words in Machine Translations

Priyesh Jain priyeshjbp@gmail.com
Sunita Sarawagi sunita@iitb.ac.in
Tushar Tomar tomartushar@cse.iitb.ac.in
Indian Institute of Technology Bombay, India

Abstract

Word-level quality scores on input source sentences can provide useful feedback to an end-user when translating into an unfamiliar target language. Recent approaches either require training custom models on synthetic data or repeatedly invoking the translation model. We propose a simple approach based on comparing probabilities from two language models. The basic premise of our method is to reason how well each source word is explained by the generated translation as against the preceding source language words. Our approach provides between 2.2 and 27.1 higher F1 score and is significantly faster than state of the art methods on three language pairs. Also, our method does not require training any new model. We release a public dataset on word omissions and mistranslations on a new language pair. [1]

1 Introduction

Neural Machine Translations exhibit human-like fluency and accuracy over many language pairs (Läubli et al., 2018), and are increasingly getting embedded in several end-user applications. A hindrance to their deployment in safety-critical applications such as healthcare and diplomacy, is that unlike human translators, a machine translation system does not provide reliable feedback when parts of a source sentence are misrepresented in the generated translation. Recent studies have discovered that NMT sentences while fluent are often inconsistent with the source (Maynez et al., 2020; Wang and Sennrich, 2020). On the target side, calibrated confidence with the generated output is a standard way to give quality feedback to the user (Kumar and Sarawagi, 2019; Lu et al., 2022). However, when a user is unfamiliar with the target language, word-level confidence in the generated translation is not actionable. We instead propose that words in *the source sentence* be assigned indicators of the loss of fidelity, so end-users could reformulate the input for possibly better translation, much like the way a human translator could ask for clarification on misunderstood words.

As a step in this direction, we attempt to solve the following problem: Given a source sentence $\mathbf{x}$ with words $x_1, \ldots x_n$ and its translation $\hat{\mathbf{y}}$ from a translation model, assign each word x_i in $\mathbf{x}$ a score $q(x_i)$ that when low indicates that the word x_i is either omitted or mistranslated in $\hat{\mathbf{y}}$. Figure 1 shows an example with a source English sentence and its translation to German. The words in the source sentence that are omitted or mis-translated are assigned relatively small scores.

Two categories of approaches have been explored so far. The first category (Zhou et al., 2021) train a supervised model to score words using synthetically inserted errors in parallel translation datasets. Vamvas and Sennrich (2022) show that such synthetic training fails to identify real cases of omissions. They instead propose a contrastive conditioning approach on the premise that the likelihood of the target $\hat{\mathbf{y}}$ will be largest when the source is stripped off the words not well translated in $\hat{\mathbf{y}}$. In addition to requiring that the stripped sentence be well-formed (enforced in the paper via a constituent parser), we show that their scoring is not as good at detecting omission and mistranslations compared to a more direct approach that we propose here.

Our contributions We propose a new approach called Target-lift for assigning quality scores to source words. The basic premise of our approach is to reason how well the generated translation $\hat{\mathbf{y}}$ explains each source word in $\mathbf{x}$. We implement this reasoning by simply comparing the conditional probability of the source word from a reverse translation model, with the unconditional probability from a source language model. Both of these are off-the-shelf models that do not require any cus-

[1] Our code is available at this link https://github.com/jain-priyesh/target-lift.git

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10683 - 10691
December 7-11, 2022

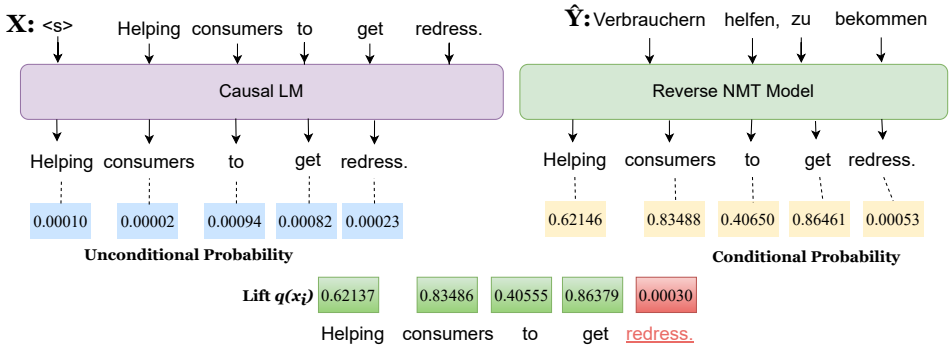

Figure 1: Quality scores on source English sentence $\mathbf{x}$ and a German translation $\hat{\mathbf{y}}$ where the word *redress* is dropped. The quality scores are very low (0.00030) for the dropped word.

tom training. Experiments on three tasks show that our scheme Target-lift achieves between 2.2 and 27.1 higher F1 score than state-of-the-art methods, while being significantly faster. Our experiments are over two existing benchmarks (En-De, Zh-En) and a new language pair (En-Hi) that we release in this paper. We are also a factor of 4 to 8 faster than the best existing method.

2 Related Work

Word-level quality metrics Recently several studies have reported the propensity of modern NMT models to generate outputs that are fluent but inconsistent with the input (Maynez et al., 2020; Martindale et al., 2019; Wang and Sennrich, 2020). This has raised interest in a more nuanced evaluation of translation outputs than standard sentence-level evaluation metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). Lommel et al. (2014) developed a Multidimensional Quality Metrics(MQM) framework for evaluation which includes word-level metrics for omission and mistranslation of source words and addition of hallucinated words in the target. Freitag et al. (2021) provides word-level annotations of such errors on the Chinese-English and English-German machine translation dataset of WMT 2020 news-translation tasks (Barrault et al., 2020). Several studies (Specia et al., 2020; Fonseca et al., 2019; Specia et al., 2018) observe omission of source words as being the most frequent type of error (Zheng et al., 2018).

Detecting Word-level Errors Post-facto estimation of the quality of translation systems is a long-standing area of interest (Kim et al., 2017) but only recently has the interest shifted to detecting word-level errors in the source or target. Some prior work have proposed to detect word-level errors based on comparison with a gold reference sentence (Kong et al., 2019; Li et al., 2021). Our focus in this work is reference-free approaches where a user is informed, during translation, of possibly mishandled source words in the target. Two categories of methods have been developed for this space. The first category (Zhou et al., 2021; Tuan et al., 2021) train special models to score words by synthetically inserting errors in a gold parallel dataset.

The second category recently proposed by Vamvas and Sennrich (2022) assign omission scores by deleting each omission candidate in the source $\mathbf{x}$ and evaluating the probability of the target ($\hat{\mathbf{y}}$) on each partial source. If the probability of the target is higher when conditioned on a partial source than the full source, the corresponding deleted words are marked as omissions. This method, called Contrastive Conditioning, has been shown in Vamvas and Sennrich (2022) to be more effective on real datasets than the first category of approaches trained on synthetic data. However, the approach of (Vamvas and Sennrich, 2022) requires as many invocations of the translation model as the number of deletion candidates. Our proposed method also falls in the second category but is significantly faster and more accurate than all existing methods.

3 Our approach

The key idea of our approach for detecting words in a source sentence $\mathbf{x}$ incorrectly handled in the generated translation $\hat{\mathbf{y}}$, is to check how well a reverse translation model with $\hat{\mathbf{y}}$ as input, explains each x_i in $\mathbf{x}$. We assume the availability of a standard auto-regressive encoder-decoder NMT model that can translate sentences from the target language to the source language. With the recent popularity of many-to-many multilingual translation mod-

Task	Method	Detection of omissions			Inference
		Precision	Recall	F1	Time (ms)
En-De	Supervised baseline	40.3 ± 5.2	6.1 ± 0.1	10.6 ± 0.2	25
	Contrastive conditioning	22.3	18.8	20.4	397
	Target-lift (Our)	20.0	28.8	**23.6**	80
Zh-En	Supervised baseline	49.6 ± 0.6	9.4 ± 1.0	15.9 ± 1.4	25
	Contrastive conditioning	25.6	62.3	36.3	750
	Target-lift(Our)	25.9	75.0	**38.5**	81
En-Hi	Contrastive conditioning	75.5	10.8	18.9	390
	Target-lift(Our)	38.1	58.1	**46.0**	90

Table 1: Omission error detection accuracy on English-German and Chinese-English dataset (sentence-level), and English-Hindi (word-level). The standard deviation of the supervised method is due to averaging over three models trained on synthetic data with different random seeds. Last column shows inference time in milliseconds. Whole dataset is evaluated on Nvidia RTX A6000 GPU, the total time is divided by the number of instances. **In all cases our method is significantly more accurate than existing methods while being a factor of 4–9 faster than the best existing method.**

els such bi-directional translations are often easily available (Aharoni et al., 2019). Using a standard auto-regressive encode-decoder architecture, the probability of each word x_i in the source $\mathbf{x}$ can be obtained from the reverse model as:

$$\Pr(\mathbf{x}|\hat{\mathbf{y}}) = \prod_{i=1}^{n} \Pr(x_i|\hat{\mathbf{y}}, \mathbf{x}_{<i}) \quad (1)$$

A simple baseline is to use $\Pr(x_i|\hat{\mathbf{y}}, \mathbf{x}_{<i})$ as a measure of the quality of word x_i. However, this score does not adequately indicate if $\hat{\mathbf{y}}$ contains a valid translation of x_i since probability $\Pr(x_i|\hat{\mathbf{y}}, \mathbf{x}_{<i})$ could be high just because x_i is a frequent next word after $x_1, \ldots, x_{i-1}$ in the source language, irrespective of $\hat{\mathbf{y}}$. We propose a simple method of disentangling this dependence: compare the conditional probability $\Pr(x_i|\hat{\mathbf{y}}, \mathbf{x}_{<i})$ with the unconditional language model probability $\Pr(x_i|\mathbf{x}_{<i})$. Accordingly, we define the score of each word x_i as the lift that the conditional model offers beyond the unconditional model as:

$$q(x_i) = \Pr(x_i|\hat{\mathbf{y}}, \mathbf{x}_{<i}) - \Pr(x_i|\mathbf{x}_{<i}) \quad (2)$$

When the lift $q(x_i)$ is high $\hat{\mathbf{y}}$ likely contains a correct translation of x_i since the unconditional LM does not support x_i. When $q(x_i)$ is low, the conditional model finds x_i unlikely, indicating that $\hat{\mathbf{y}}$ has omitted or mistranslated x_i. Thus, by simply thresholding on the $q(x_i)$ scores we can identify words that are poorly represented in a generated translation $\hat{\mathbf{y}}$. We call this method Target-lift. When a word is segmented into subwords, we take the first subword as the probability of the word. Figure 1 presents an example.

The unconditional probability $\Pr(x_i|\mathbf{x}_{<i})$ can be obtained from any off-the-shelf causal language model. However, since the models used to compute the two probabilities could in general be uncomparable, we recalibrate the probabilities $\Pr(x_i|\mathbf{x}_{<i})$ from the unconditional model. We use the temperature scaling method (Platt, 1999) where the temperature is chosen based on a validation dataset. Temperature scaling changes the skewness of a probability distribution $\Pr(x_i)$ as follows: $P_T(x_i) \propto \Pr(x_i)^{\frac{1}{T}}$. Large values of T causes the distribution to be less skewed and has been found by several previous studies to correct the ill-effects of over-fitting in modern deep networks (Kumar and Sarawagi, 2019).

Empirical evaluation show that this simple score provides significant gains over existing methods while relying entirely on off-the-shelf models. The running time is also significantly lower since we require only two invocations of sequence models unlike (Vamvas and Sennrich, 2022).

4 Experiments

We evaluate our method on three datasets and compare with two existing methods.

Datasets We used English-German and Chinese-English machine translations from the WMT 2020 news translation task Barrault et al. (2020) along with annotations made by Freitag et al. (2021). The train and test splits are exactly the same as in Vamvas and Sennrich (2022).We introduce a new English-Hindi dataset of error tagged source words on 2000 sentences chosen from the IITB

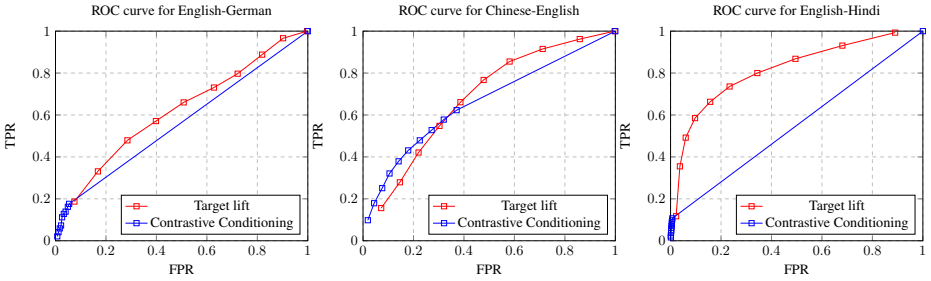

Figure 2: The ROC curves on testset of the language pairs English-German, Chinese-English and English-Hindi where the unconditional probabilities are re-calibrated with temperatures 3.5, 4.0 and 1.5 respectively.

English-Hindi Parallel corpus (Kunchukuttan et al., 2018) and translated using Marian NMT model (Junczys-Dowmunt et al., 2018). More details of the sentence selection and human annotation process appear in the Appendix. A summary of the statistics of all three datasets appear in Table 5 in the Appendix.

Methods compared We compare our method Target-lift against these two recent methods.
Supervised baseline:(Zhou et al., 2021; Tuan et al., 2021) that assigns word-scores using a custom model trained on synthetic data. We reproduce the numbers from Vamvas and Sennrich (2022) that implements these models following the predictor-estimator model in (Kim et al., 2017).
Contrastive Conditioning:(Vamvas and Sennrich, 2022) We use the author's implementation.

Setup We use mBART50, a seq-to-seq Transformer pre-trained (Tang et al., 2021) on 50 languages and fine-tuned for multilingual machine translation. The one-to-many[2] variant is used if English is the source language and the many-to-one[3] variant if English is the target. We use a mBART50 model[4] pretrained for the causal language modelling task to compute unconditional probability. Before softmax, we scale the logits vector of the causal model using a temperature selected using the validation set of that pair.

4.1 Overall Results

Table 1 presents the overall results. For En-De and Zh-En we report sentence-level F1 scores of omission detection following earlier work. For En-Hi we report word-level errors since *82%* test sentences in the dataset contains word errors unlike the other two where less than 16% have errors. Observe that on all three datasets our method provides significantly higher F1 than existing methods. On En-Hi, where we measure word-level F1, the differences are particularly striking. Figure 2 shows the ROC curves for all the language pairs and shows that Target-lift dominates Contrastive conditioning across different values of thresholds. Table 1 shows that in terms of running time Target-lift that scores all words in two calls to language models, is significantly faster than Contrastive Conditioning, that requires as many calls to the translation model as the number of omission source candidates. In the Appendix we present anecdotes to further demonstrate the working of Target-lift.

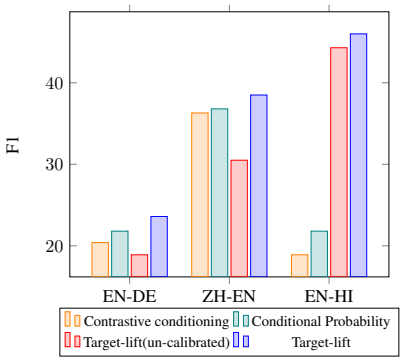

Figure 3: F1 with different variants of our method compared against SOTA Contrastive Conditioning method. Variants are: (1) Only conditional probabilities from a reverse translation model and (2) Uncalibrated version of our method (Section 4.2).

4.2 Ablation study

Effect of Calibration We also compare our method with ablations on the different components of its design. (1) Scoring purely based on reverse conditional probability $\Pr(x_i|\hat{\mathbf{y}}, \mathbf{x}_{<i})$ and

[2] https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt
[3] https://huggingface.co/facebook/mbart-large-50-many-to-one-mmt
[4] https://huggingface.co/docs/transformers/v4.19.4/en/model_doc/mbart#transformers.MBartForCausalLM

(2) Not calibrating the unconditional probability. Figure 3 shows that both variants are much worse than Target-lift. Interestingly, in all cases the Contrastive Conditioning method is even worse than simply scoring on reverse conditional probability.

Method of Aggregating subword probabilities We compare different subword aggregation methods: (1) First subword probability (default), (2) Product of subword probabilities, and (3) Minimium of subword probabilities in Table 2. We find that our approach of using the First subword provides best performance overall, possibly because the probability of subsequent subwords are often strongly dependent on the first subword.

Language Pair	Aggregation Method	Precision	Recall	F1
En-De	First	**24.7**	**46.8**	**32.4**
	Product	21.1	43.6	28.5
	Minimum	21.8	46.8	29.7
En-Hi	First	**22.5**	41.2	**29.1**
	Product	21.8	**43.1**	29.0
	Minimum	22.2	42.0	29.0

Table 2: Precision Recall and F1 values of language pairs English-German and English-Hindi on Devset for different subword probabilities aggregation methods. For Chinese there was no need for subword aggregation since it was a character-level model.

Language Pair	Comparison Method	Precision	Recall	F1
En-De	Difference	**24.7**	**46.8**	**32.4**
	Relative difference	24.7	39.4	30.3
	Ratio of Probabilities	19.9	42.6	27.1
En-Zh	Difference	27.9	**70.3**	**39.9**
	Relative difference	27.6	66.0	38.9
	Ratio of Probabilities	**28.2**	64.9	39.3

Table 3: Precision Recall and F1 values of language pairs English-German and Chinese-English on Devset for different methods of comparing probabilities.

Method of comparing probabilities In Table 3 we evaluate alternative methods of comparing the conditional and unconditional probabilities: difference (default), ratio, relative difference. We found the best results with the difference of probabilities as used in Equation 2.

Masking multiple words One concern we had with Target-lift was that if a word x_i is strongly dependent on x_{i-1} then both conditional and unconditional probabilities will be similar irrespective of the quality of translation. To address this, we explored another variant where we masked word x_{i-1} too when calculating both conditional and unconditional probabilities. Comparing results of this method in Table 4 with those of our default in Table 1 show that simple masking of the preceding words does not work, and more evolved strategies may need to be explored in the future.

Language Pairs	Precision	Recall	F1
En-De	9.6	75.3	17.1
Zh-En	18.0	81.0	29.5
En-Hi	18.8	69.1	29.6

Table 4: Precision Recall and F1 of all language pairs on Testset when x_{i-1} word is also masked while calculating both conditional and unconditional probabilities.

4.3 Discussion of Limitations

Our method could mishandle idiomatic phrases and we show one example in Figure 4. Although "lucky escape" is correctly translated Target-lift still tags it but not "lucky". Another limitation of any word-based method is that subtle mistakes in placement of prepositions and other connectives may be missed. In example 2 of Figure 4, the sense of "to set [] in place" is missed.

(1) Source: A puppy had a lucky escape after fire crews were called to lift her to safety when she somehow got herself stuck 50ft up on a precarious cliff ledge .
Translation: एक पिल्ला उस समय बाल-बाल बच गया जब उसे सुरक्षित निकालने के लिए दमकल कर्मियों को बुलाया गया और वह किसी तरह 50 फीट की ऊंचाई पर एक खतरनाक चट्टान पर फंस गई।
(2) Source: Corrective measures are needed to set the faulty planning in place .
Translation: गलत योजना निश्चित करने के लिए सही उपाय की ज़रूरत होती है ।

Figure 4: English-Hindi sentence pairs where Target-lift fails on some cases.

5 Conclusions and Limitations

We presented a simple method of assigning quality scores to source words in a translation model as the lift of the conditional probability from a reverse translation model, over the unconditional probability from the language model. In spite of its simplicity, our method is more accurate and faster than state-of-the-art methods, and does not require training any special models. We also release a new dataset of word-level omissions and mistranslations on a new language pair.

Acknowledgement We thank Keshav Agarwal for helping create the English-Hindi dataset.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2019. Neural machine translation with adequacy-oriented learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6618–6625.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *CoRR*, abs/1903.00802.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Guanlin Li, Lemao Liu, Conghui Zhu, Rui Wang, Tiejun Zhao, and Shuming Shi. 2021. Detecting source contextual barriers for understanding neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3158–3169.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation. In *ACL*.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. 2021. Quality estimation without human-labeled data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lili Mou, Xinyu Dai, Jiajun Chen, and Zhaopeng Tu. 2018. Modeling past and future for neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:145–157.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

6 Appendix

Dataset split	Number of segments	
	Total	With an omission error
En-De Dev	1418	187
En-De Test	4839	484
Zh-En Dev	1999	516
Zh-En Test	8851	1569
En-Hi Dev	500	350
En-Hi Test	1500	1233

Table 5: Statistics for the three datasets used in our experiments. The validation-test splits for the first two datasets are the same as in (Vamvas and Sennrich, 2022). The English-Hindi dataset is introduced by us in this paper.

English-Hindi Dataset Annotations From the En-Hi IITB parallel corpus we filtered 2000 sentences according to length (8-24 words) and number of un-aligned words (5 and more) over alignments computed using Awesome align Dou and Neubig (2021). The selected sentence pairs along with generated translation(using Marian NMT) are given to two human annotators. They tagged omitted and mistranslated words as erroneous. The first 500 sentences are set as validation set and rest used as test set.

6.1 Examples of Errors predicted by Target-lift

English-German Examples

Source: He added, "It's backfired now **though,** that's the sad thing."
Translation: Er fügte hinzu: „Es ist jetzt nach hinten losgegangen, aber das ist das Traurige“.
Target lift predicted error token: though
Contrastive conditioning predictions: None
Annotator rating: Mistranslation
*The word is not omitted but, is incorrectly used. Possible correct translation-"*Er fügte hinzu: “Es ist jetzt aber nach hinten losgegangen, das ist das Traurige.”*"*

Source: Seth **Meyers** Loses It **Laughing** During Donald Trump 'Tango & Cash' **Impression**
Translation: Seth Myers bekommt Lachanfall bei „Tango und Cash“-Auftritt von Donald Trump
Target lift predicted error tokens: Meyers, Laughing, Impression
Contrastive conditioning predictions: None
Annotator rating: Mistranslation - Auftritt = ‘appearance of’
Here, Meyers is misspelt, ‘Loses if Laughing’ and ‘Impression’ are incorrectly translated

Source:News outlets report 39-year-old Jerrontae Cain was sentenced Thursday on charges including **being a felon** in possession of a gun in the 2017 attack on 42-year-old Nicole Gordon.
Translation: Nachrichtenagenturen berichten, dass der 39-jährige Jerrontae Cain am Donnerstag unter anderem wegen des Besitzes einer Waffe bei dem Angriff 2017 auf die 42-jährige Nicole Gordon verurteilt wurde.
Target lift predicted error token: ‘felon’
Contrastive conditioning predictions: *None*
Annotator rating: Omission
Here ‘being a felon’ is not correctly translated. That is why ‘felon’ is tagged as omitted/mistranslated.

Chinese-English Examples

Source: 国有企业和优势民营企业走进赣南革命老区。
Translation: State-owned enterprises and dominant private enterprises visited the Gannan revolutionary base area.
Target lift predicted error token: 老 =’old’
Contrastive conditioning predictions: None
Annotator rating: Omission

Source: 中新网北京9月27日电 (记者 杜燕)为加强节前市场监管执法，北京市市场监管局在国庆节前夕检查各类经营主体2000余户。
Translation: Beijing, Sept. 27 / prnewswire-asianet / -- in order to strengthen market supervision and law enforcement before the festival, the Beijing Municipal Market Supervision Bureau inspected more than 2,000 households of various business entities on the eve of the National Day.
Target lift predicted error token: 中新网，日电，杜，燕 = ‘China News Service’, ‘NEC’,’Du’, ‘Yan’
Contrastive conditioning predictions: *None*
Annotator rating: Omission - (Report Du Yan) - Beijing
Here, the Contrastive conditioning method doesn’t flag errors, although “China News Service” and “Report Du Yan” are clearly omitted, which is caught by our method.

Source:协议规定，伊朗只能用第一代IR-1型离心机来提炼铀。
Translation: According to the agreement, Iran can only use the first generation IR-1 centrifuge to refine uranium.
Target lift predicted error token: No errors
Constrastive conditioning predictions: 机来 = ‘machine’
Annotator rating: No error

English-Hindi Examples

Source: Beneath is a **rather clumsily rendered** bull with head **turned up** .
Translation: इसके नीचे सिर झुका हुआ एक सांड होता है।
Target lift predicted error token: rather, clumsily, rendered, turned, .
Contrastive conditioning predictions: None
True error tokens: rather, clumsily, rendered, turned, up
Almost all the tokens were predicted and contrastive conditioning failed to predict even one.

Source: This is a little structure built of huge **almost boulder-like** blocks of stones .
Translation: यह एक छोटी सी संरचना है जो लगभग पत्थर जैसे पत्थरों के बड़े-बड़े खंडों से बनी है।
Predicted error tokens: built, huge, almost, boulder-like, .
Contrastive conditioning predictions: None
True error tokens: almost, boulder-like
Same case as above

Source: The music of the 'poongi ' has a **sinuous** quality , which makes a dancer swirl and dance like a serpent .
Translation: 'पूंगी' के संगीत में एक शिथिल गुणवत्ता है, जो एक नर्तक को सर्प की तरह घूमता और नृत्य करता है।
Target lift predicted error token: sinuous, .
Contrastive conditioning predictions: None
True error tokens: sinuous

Source: Two buffalo horns tied together with two reed pipes at the narrower end with four finger holes .
Translation: भैंस के दो सींग चार उंगलियों के छेद के साथ संकरे छोर पर दो रीड पाइपों से बंधे हुए थे।
Target lift predicted error token: '.' *(period)*
True error tokens: No-error
Contrastive conditioning predictions: None
Failure case of our method, '.' is flagged as an error. This might be due to tokenised source sentence.