On the Limits of Minimal Pairs in Contrastive Evaluation

Jannis Vamvas[1] and **Rico Sennrich**[1,2]

[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh
`{vamvas,sennrich}@cl.uzh.ch`

Abstract

Minimal sentence pairs are frequently used to analyze the behavior of language models. It is often assumed that model behavior on contrastive pairs is predictive of model behavior at large. We argue that two conditions are necessary for this assumption to hold: First, a tested hypothesis should be well-motivated, since experiments show that contrastive evaluation can lead to false positives. Secondly, test data should be chosen such as to minimize distributional discrepancy between evaluation time and deployment time. For a good approximation of deployment-time decoding, we recommend that minimal pairs are created based on machine-generated text, as opposed to human-written references. We present a contrastive evaluation suite for English–German MT that implements this recommendation.[1]

1 Introduction

Contrastive evaluation is one of the most widely used evaluation techniques for generative language models, both for causal models (Linzen et al., 2016) and sequence-to-sequence models (Sennrich, 2017). Various phenomena have been analyzed using this technique, including syntax (Marvin and Linzen 2018; among others), word sense disambiguation (Rios et al. 2017; among others), document coherence (Bawden et al. 2018; Beyer et al. 2021; among others), and grammatical acceptability in general (Warstadt et al., 2020; Xiang et al., 2021).

Contrastive evaluation allows for a targeted, automated evaluation of generative models but is restricted to a specific behavioral interface, namely the ranking of pre-defined *minimal pairs*. However, most models in application areas such as translation or conversation are deployed to produce 1-best sequences, exposing a different behavioral interface to users. While this limitation of contrastive

evaluation is well known, its practical relevance has been unclear.

We show that under certain conditions, the gap between evaluation and deployment can indeed cause misleading results. As a main factor we identify the *distributional discrepancy* of contrastive evaluation datasets: Minimal pairs are usually derived from human-written references, but when deployed, a model is conditioned on its own output.

To measure the effect of this factor on evaluation, we focus on neural machine translation (NMT) systems. Our approach is to test *implausible research hypotheses* in addition to plausible ones. We find that distributional discrepancy increases the number of false positives regarding implausible hypotheses. They particularly occur when evaluating distilled NMT models (Kim and Rush, 2016), indicating that in such models, ranking behavior on noisy sequences diverges from generative behavior.

We also propose a way to reduce the distributional discrepancy of minimal pairs. Our experiments show that false positives can be largely avoided by using machine-generated text instead of human-written text. This inspires us to release **DistilLingEval**, a variant of the LingEval97 English–German MT evaluation suite (Sennrich, 2017) that uses MT-generated references.

We recommend that future efforts to create contrastive datasets for the evaluation of language generation models minimize distributional discrepancy between evaluation and deployment. Due to the possibility of false positives, linguistic conclusions about *knowledge* or *abilities* of models should be corroborated by additional evidence from a more natural setting.

2 Background and Related Work

2.1 Contrastive Evaluation

Contrastive evaluation compares the probability scores that a model assigns to two minimally dif-

---

[1]https://github.com/ZurichNLP/distil-lingeval

ferent sequences. For example, the sentences *"The cats sleep"* and *"The cats sleeps"* differ in verb number only; if a model assigns higher scores to sentences of the first kind than to sentences of the second kind, it is said to prefer verb forms in agreement with the noun (Linzen et al., 2016).

An established method for scoring is to compute the score for a full sentence $X = x_0, x_1, ..., x_n$ as the sum of token log-probabilities predicted by the model θ (Marvin and Linzen, 2018):

$$\text{score}(X) = \sum_{i=0}^{n} \log p_\theta(x_i|x_{<i}) \qquad (1)$$

When contrastive evaluation is applied to sequence-to-sequence models, two target sequences are scored given the same source sequence X (Sennrich, 2017). We follow previous work and normalize sequence-to-sequence scores by the length of the target sequence Y:

$$\text{score}(Y|X) = \frac{1}{|Y|} \sum_{i=0}^{|Y|} \log p_\theta(y_i|X, y_{<i}) \quad (2)$$

2.2 Limitations of Forced Choice

Since a limited set of variants is scored, contrastive evaluation presents the model with a forced choice. In fact, scoring a pre-defined sequence is related to *teacher forcing*, i.e., the conditioning of a model on a ground truth prefix during training. Whenever an application involves unconstrained generation, a discrepancy between evaluation and deployment arises that is comparable to the *exposure bias* of cross-entropy training (Ranzato et al., 2016).

With regard to syntactic evaluation of language models, Newman et al. (2021) point out that contrastive evaluation and evaluation of systematicity across a paradigm do not necessarily describe a model's likely behavior. They propose to analyze the complete search space, which, however, is difficult to implement in many use cases. We pursue a different strategy and create minimal pairs that are more similar to sequences the model will likely generate at deployment time.

3 Experiments

In previous work, contrastive evaluation has commonly been used to test *plausible* research hypotheses, for example the hypothesis that RNNs can predict long-distance number agreement (Gulordava et al., 2018), or the hypothesis that word dropout

improves pronoun resolution in translation (Fernandes et al., 2021). In this paper, we are interested in *implausible* hypotheses and in how the testing of such hypotheses is affected by the limitations described in the previous section.

We formulate two implausible hypotheses about NMT systems, which we mark with an asterisk (*):

1. **Vague language:*
   NMT systems make liberal use of vague placeholder words. Specifically, English–German models use the German placeholder noun *Ding* ('thing') ubiquitously.

2. **Hypercorrection:*
   NMT systems have a tendency for hypercorrect language. Specifically, English–German models tend to use genitive case with prepositions that require dative case.

Examples are given in the next section. The two hypotheses are chosen because they seem implausible both theoretically and empirically. From a theoretical standpoint, both linguistic phenomena rarely occur in the training data and the model is unlikely to adopt them broadly. Furthermore, the cognitive and social factors that cause the phenomena in human speech do not apply to neural language models. Empirically, we find that both phenomena are indeed very rare in neural machine translations, independent of model quality.

For comparison, we also test two plausible hypotheses about NMT systems:

3. *Polarity affix deletion:*
   NMT systems sometimes omit negation affixes, changing the polarity of a word (Hossain et al., 2020). Specifically, English–German models sometimes omit the negation prefix *un-* from German words (Sennrich, 2017).

4. *Clause omission:*
   NMT systems sometimes omit a clause from the translated sentence (Tu et al., 2016).

3.1 Test Set Creation

For each of the four hypotheses, we create an English–German contrastive test set. For *vague language*, *polarity affix deletion* and *clause omission*, we use the newstest datasets 2009–2016 as a data source. For *hypercorrection*, we combine five data sources: newstest 2009–2019 as well as OpenSubtitles2016 (Lison and Tiedemann,

2016), TED2020 (Reimers and Gurevych, 2020), QED (Abdelali et al., 2014) and JW300 (Agić and Vulić, 2019), which are provided by OPUS (Tiedemann, 2012).

Vague language Contrastive variants are created by replacing a random noun in each reference with an uninflected *Ding* 'thing', which is a common replacement noun in spoken German (Vogel, 2020):

English: *Prague Stock Market falls to minus by the **end of the trading day***
German (correct): *Die Prager Börse stürzt gegen **Geschäftsschluss** ins Minus.*
German (contrastive): *Die Prager Börse stürzt gegen **Ding** ins Minus.*

Hypercorrection To create contrastive variants for hypercorrect genitives, we select references containing German propositions that require dative in Standard German, but are sometimes used hypercorrectly with a genitive case (Hentschel and Weydt, 2013).[2] We construct contrastive variants by converting the dative case into genitive case:

English: *I've loved you ever **since that day** in the rose garden.*
German (correct): *Ich liebe dich **seit dem Tag** im Rosengarten.*
German (contrastive): *Ich liebe dich **seit des Tags** im Rosengarten.*

Polarity affix deletion Contrastive variants are created by deleting the prefix *un-* from adjectives, adverbs and nouns in the German references in cases where this changes the polarity of the word, similar to the test set created by Sennrich (2017):

English: *The probes **unexpectedly** become faster or slower.*
German (correct): *Die Sonden werden **unerwartet** schneller oder langsamer.*
German (contrastive): *Die Sonden werden **erwartet** schneller oder langsamer.*

Clause omission Contrastive variants are created by deleting a clause from the reference. As clauses we treat token sequences segmented by the Stanza sentence splitter (Qi et al., 2020):

English: *And even if it could be proved for humans - **how would one want to prove it for rats?***

German (correct): *Und selbst wenn man das für den Menschen beweisen könnte: **Wie wollte man es bei Ratten nachweisen?***
German (contrastive): *Und selbst wenn man das für den Menschen beweisen könnte:*

3.2 Human-Written References

The above test sets are derived from naturally occurring parallel text, which is common practice when creating contrastive datasets for MT (Sennrich, 2017; Rios et al., 2017; Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019; Raganato et al., 2019; Sugiyama and Yoshinaga, 2019; Nagata and Morishita, 2020; Shimazu et al., 2020; Lopes et al., 2020; He et al., 2020; Stojanovski et al., 2020). However, comparisons have shown that human-written references are different from machine translations in that they contain more noise and have more linguistic diversity (Zhang et al., 2018; Vanmassenhove et al., 2019).

We propose to measure the "distance" between a pre-defined target sequence and the 1-best translation $\hat{Y}$ generated by an MT system as the difference in log-scores (according to Equation 2) that the system assigns to the two sequences. Furthermore, we define the *distributional discrepancy* of a contrastive evaluation dataset as the mean difference in scores between the 1-best translation and the preferred variant:

$$\text{score\_preferred} = \max(\text{score}(Y^{correct}), \text{score}(Y^{contrast.}))$$

$$\text{discrepancy} = \frac{1}{n} \sum_{i=0}^{n} \text{score}(\hat{Y}_i) - \text{score\_preferred}_i$$

It should be noted that this definition of distributional discrepancy is mainly useful for comparing multiple test sets with respect to a single model. It is less useful for assessing a single test set with respect to multiple models, because score differences are not necessarily comparable between models.

3.3 Machine-Generated References

With the goal of reducing distributional discrepancy, we create versions of our test sets that use machine-generated references. First, we re-translate the sources from our test sets using commercial NMT systems.[3] We then repeat the steps described in Section 3.1 to create contrastive variants.

---

[2]We use the prepositions *entgegen, entsprechend, gegenüber, gemäß, nahe, nebst, (mit)samt* and *seit*.

[3]We used *Amazon Translate, DeepL Translator, Google Translate*, and *Microsoft Translator* for 25% sentences each.

Hypothesis	Test set type	Discrepancy of test set		Reported accuracy	
		TRANSFORMER	DISTILLED	TRANSFORMER	DISTILLED
Vague language	human references	1.2 ± 0.0	2.5 ± 0.1	99.1 ± 0.1	94.7 ± 0.4
	machine references	0.3 ± 0.0	0.7 ± 0.0	99.9 ± 0.0	98.7 ± 0.2
Hypercorrection	human references	1.3 ± 0.0	2.7 ± 0.1	95.4 ± 0.3	91.2 ± 0.5
	machine references	0.4 ± 0.0	1.1 ± 0.1	99.9 ± 0.1	99.6 ± 0.4
Polarity affix del.	human references	1.3 ± 0.0	2.7 ± 0.1	94.0 ± 1.1	78.3 ± 0.9
	machine references	0.3 ± 0.0	0.7 ± 0.1	96.7 ± 1.5	93.9 ± 1.1
Clause omission	human references	1.3 ± 0.0	2.8 ± 0.1	75.5 ± 3.7	71.3 ± 0.7
	machine references	0.3 ± 0.0	0.7 ± 0.0	87.7 ± 2.7	86.3 ± 2.7

Table 1: Results for four different hypotheses about English–German NMT systems. An asterisk (*) marks hypotheses that are a priori implausible. The table reports distributional discrepancies of different test set types, as well as the accuracy scores achieved by non-distilled and distilled systems when evaluated with the test sets. We report averages and standard deviations across three models trained independently with different random seeds.

Validation Since some machine-generated references contain errors, a validation step is needed. The validation should ensure that (a) the machine references are correct with respect to the linguistic phenomenon at hand, and that (b) no undesired bias is introduced into the evaluation.

We use a semi-automatic approach and look for lexical overlap with the human references regarding the phenomenon. For example, in the case of polarity affix deletion, we label the machine reference as correct if it contains the same polarity word as the human reference. Otherwise we manually check whether the machine reference might be incorrect, but only if it contains the same polarity word as the human contrastive variant. This occurs rarely, and most of the time we find that it is the original human reference that is incorrect while the machine reference is correct. In the rare cases where the machine reference is verifiably incorrect with regard to the phenomenon, we use it as the contrastive variant and derive the correct variant manually.

Machine references that have no phenomenon-specific lexical overlap to the human references are dropped from the test set because they cannot be automatically validated. This raises the question whether test sets created in such a way contain undesired bias.

Dataset Bias We discuss two kinds of bias that might be introduced. First of all, by only including machine references that can be classified automatically as either correct or incorrect based on the human references, the distribution of the machine-generated test set could become more similar to the human-written test set. However, our experiments show that the difference in distributional discrepancy between the two test sets is sufficiently large. Future work could avoid this bias by employing human annotators to validate machine references.

Secondly, it might be that machine references only use the phenomenon in unambiguous contexts. This would cut off the long tail of human-written test samples that is especially challenging for NLP models. While such a bias is likely to be introduced to a degree, we see it is a *desired* bias, since our goal is to reduce distributional discrepancy between a test set and the generative behavior of an evaluated system.

3.4 Experimental Setting

We evaluate two types of NMT systems:

1. TRANSFORMER: Transformer models of size 'big' (Vaswani et al., 2017).
2. DISTILLED: Transformer models of size 'small' distilled from (1) using sequence-level knowledge distillation (Kim and Rush, 2016).

Training For both types, we trained three models with different random seeds. To train the TRANSFORMER models, we used similar data and configuration as Ng et al. (2019), using Fairseq (Ott et al., 2019). We used the English–German parallel training data from the WMT19 news translation task (Barrault et al., 2019). Sentences longer than 250 tokens and pairs with a length ratio larger than 1.5 were filtered, resulting in 42.9M sentence pairs used for training and distillation. We selected the

best checkpoint with respect to BLEU based on the *newstest* sets from the preceding years.

Distillation We then used each of the three TRANSFORMER models as a teacher to train an individual student model. A comparison of hyperparameters is provided in Appendix A.

For decoding we always use beam search with size 5.

Model Quality The models of type TRANSFORMER achieve an average BLEU score of 37.3 ± 0.3, while the DISTILLED models achieve 35.7 ± 0.4 BLEU when evaluated on newstest19.

3.5 Results

The left-hand side of Table 1 shows the distributional discrepancies of the test sets. As expected, the test sets derived from human-written references have a higher discrepancy, while those derived from machine-generated references are closer to what the evaluated model would generate.

The right-hand side of Table 1 shows the reported *accuracy* of the evaluated models, i.e. the ratio of test instances where the model prefers the correct variant over the contrastive variant. While all the accuracies are much better than random, the results for implausible hypotheses seem to indicate that models do occasionally generate the implausible phenomena, and that distilled models generate them more often than other models. Since this is not reflected by the actual generative behavior, the testing of implausible hypotheses shows the danger of false positives.

The test sets with machine-generated references produce far fewer false positives. The reported accuracy is higher with machine references also for the plausible hypotheses, but a gap to 100% accuracy remains, which is in line with previous work on these types of NMT errors (Hossain et al., 2020; Tang et al., 2021; Tu et al., 2016).

4 Dataset Release

Given the improved specificity of test sets with machine-generated references, we release corresponding test sets for other phenomena in the LingEval97 test suite (Sennrich, 2017), terming our dataset variant **DistilLingEval**.

LingEval97, the original test suite, is a collection of 97k contrastive translation pairs for 13 different error types in English–German translation. Building on LingEval97, we create test sets with

Error type	Human Ref.	MT Ref.
clause_omission	1104	1025
hypercorrect_genitive	3404	635
np_agreement	24 055	10 595
placeholder_ding	18 647	18 659
polarity_affix_del	408	180
polarity_particle_kein_del	554	201
polarity_particle_nicht_del	2561	888
subj_verb_agreement	31 978	6701

Table 2: Number of samples per DistilLingEval error type. Error types with machine-generated references tend to have fewer samples, which is discussed in Section 3.3.

machine-generated references for the following error types, in addition to the ones discussed in the previous section: *noun phrase agreement*, *subject-verb agreement* and other *polarity deletion* phenomena involving the German negation lexemes *kein* and *nicht*. Results for these test sets are reported in Table 3, and further results for a state-of-the-art NMT system are provided in Appendix C. Table 2 provides an overview of the test set sizes per error type in DistilLingEval.

5 Discussion

By testing implausible hypotheses, we demonstrate the risk of drawing wrong inferences about generative behavior of (conditional) language models, especially if there is a large distributional discrepancy between minimal pairs and generated sequences.

This problem is especially apparent for distilled NMT models, which perform poorly on human-written minimal pairs because they were never exposed to such a distribution during training. While this indicates that distilled NMT models are less robust against improbable contexts, human-crafted minimal pairs also become less useful to predict their unconstrained generative behavior.

The danger of false positives from minimal pairs highlights the fact that behaviorist approaches to measuring knowledge are limited to the behavioral interface that is observed. Systematic assessments of linguistic *knowledge* or syntactic *abilities* of neural models should be qualified accordingly, in case minimal pairs are the primary analysis method. We suspect that whenever a broad range of hypotheses is tested, including phenomena that are rarely observed in actual machine-generated text, the risk of false positives is increased.

Error Type	Test set type	Discrepancy of test set		Reported accuracy	
		TRANSF.	DISTILLED	TRANSF.	DISTILLED
np_agreement	human references	2.0 ± 0.6	2.6 ± 0.1	95.9 ± 2.5	94.4 ± 0.8
	machine references	0.5 ± 0.2	0.7 ± 0.1	98.8 ± 0.8	99.0 ± 0.4
polarity_particle_kein_del	human references	1.3 ± 0.0	2.7 ± 0.1	95.3 ± 0.8	90.7 ± 0.9
	machine references	0.2 ± 0.0	0.6 ± 0.1	99.8 ± 0.3	99.8 ± 0.3
polarity_particle_nicht_del	human references	1.3 ± 0.0	2.6 ± 0.1	95.5 ± 0.2	87.9 ± 0.7
	machine references	0.3 ± 0.0	0.7 ± 0.0	99.7 ± 0.3	98.8 ± 0.1
subj_verb_agreement	human references	1.2 ± 0.0	2.6 ± 0.1	97.1 ± 0.3	91.4 ± 0.3
	machine references	0.3 ± 0.0	0.7 ± 0.0	99.2 ± 0.1	97.9 ± 0.3

Table 3: Test set discrepancies and model accuracies for the other four error types included in DistilLingEval (in addition to the four error types in Table 1).

We thus recommend that minimal pairs be constructed from machine-generated text in evaluation settings where unconstrained generation is the behavior of interest. This is relatively straightforward for sequence-to-sequence evaluation, as we demonstrated in our experiments. For other settings, e.g. the evaluation of dialogue models, obtaining useful machine-generated text might require more elaborate techniques, such as round-trip translation.

However, human-crafted minimal pairs remain valuable in other use cases. While machine-generated pairs may be more appropriate when the main interest is to study the model's behavior close to its mode, e.g. in a sequence-to-sequence task, human-written pairs (or pairs that are machine-generated to be different from the training distribution on purpose) may tell us more about the robustness of models outside the mode. For example, terminology-constrained or interactive applications depend on robustness against improbable contexts, and contrastive evaluation indicates that current NMT systems lack such robustness (Stojanovski et al., 2020). Similarly, syntactic evaluation of language models using randomly generated or nonsensical sentences (Gulordava et al., 2018; Warstadt et al., 2020) can be seen as method to assess the robustness of a model under improbable input, rather than as an assessment of generative capabilities in general.

6 Conclusion

We show that there are conditions where contrastive evaluation leads to false positives if generative behavior is inferred from behavior under forced choice. Experiments with English–German NMT indicate that the gap between the two behavioral interfaces is especially high when human-written text is used to create minimal pairs. Using machine-generated text largely reduces the gap. We recommend that human-written minimal pairs are mainly used for assessing the robustness of models, but that for predicting the generative behavior of language models, machine-generated minimal pairs are used.

Broader Impact

For language generation systems to be deployed, they should behave according to specified principles in a robust way. Typical requirements are linguistic acceptability, avoidance of undesirable societal biases (Sheng et al., 2021), and the avoidance of harmful speech acts. Contrastive evaluation is one of several methods that can help predict the behavior of language generation systems. However, to our knowledge the method has been mainly used to evaluate linguistic acceptability, and less to evaluate ethically sensitive aspects of generation.

It is crucial that evaluation methods have a high predictiveness regarding the behavior of a deployed system. On the one hand, lack of sensitivity can lead to unforeseen negative impact. On the other hand, lack of specificity – which we address in this paper – reduces the usefulness of comparisons between systems.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? Targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Elke Hentschel and Harald Weydt. 2013. *Handbuch der deutschen Grammatik*. De Gruyter.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third*

Conference on Machine Translation: Research Papers, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Masaaki Nagata and Makoto Morishita. 2020. A test set for discourse translation from Japanese to English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3704–3709, Marseille, France. European Language Resources Association.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Annette Rios, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Sho Shimazu, Sho Takase, Toshiaki Nakazawa, and Naoaki Okazaki. 2020. Evaluation dataset for zero pronoun in Japanese to English translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3630–3634, Marseille, France. European Language Resources Association.

Dario Stojanovski, Benno Krojer, Denis Peskov, and Alexander Fraser. 2020. ContraCAT: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Gongbo Tang, Philipp Rönchen, Rico Sennrich, and Joakim Nivre. 2021. Revisiting Negation in Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:740–755.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural

machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Petra M. Vogel. 2020. *Dingsbums and Thingy: Placeholders for Names in German and Other Languages*, page 362–383. Cambridge University Press.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Dakun Zhang, Josep Crego, and Jean Senellart. 2018. Analyzing knowledge distillation in neural machine translation. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 23–30.

A Hyperparameters

Name	N	d_{model}	d_{ffn}	h	Parameters
TRANSFORMER (big)	6	1024	8192	16	269.7M
DISTILLED (small)	4	512	2048	4	50.9M

Table 4: Hyperparameters of the Transformer variants used for the experiments

B Examples

Example Inputs (English–German)	Score Assigned by Model
Source: *Yesterday evening, the committee wanted to vote on the **appointment**.*	
1-best translation by the evaluated system: *Gestern Abend wollte der Ausschuss über die **Ernennung** abstimmen.*	-0.09
Minimal pair based on a human-written reference:	
– Correct: *Gestern Abend wollte das Gremium über die **Personalie** abstimmen.*	-3.61
– Incorrect: *Gestern Abend wollte das Gremium über die **Ding** abstimmen.*	-2.34
Minimal pair based on a machine-generated reference (Amazon Translate):	
– Correct: *Gestern Abend wollte der Ausschuss über die **Ernennung** abstimmen.*	-0.09
– Incorrect: *Gestern Abend wollte der Ausschuss über die **Ding** abstimmen.*	-1.25
Source: *Why did Judah lose its land **and** temple?*	
1-best translation by the evaluated system: *Warum hat Juda sein Land **und** seinen Tempel verloren?*	-0.11
Minimal pair based on a human-written reference:	
– Correct: *Warum verlor Juda sein Land **mitsamt dem** Tempel?*	-2.58
– Incorrect: *Warum verlor Juda sein Land **mitsamt des** Tempels?*	-2.55
Minimal pair based on a machine-generated reference (DeepL):	
– Correct: *Warum hat Juda sein Land **und** seinen Tempel verloren?*	-0.11
– Incorrect: N/A	

Table 5: Examples of human-written and machine-generated minimal pairs for the *Vague language* hypothesis (top) and the *Hypercorrection* hypothesis (bottom). The log-scores are computed by an NMT model of type DISTILLED.

The first example demonstrates that a model often assigns a lower score to the correct human reference than to the incorrect machine reference. The human reference differs from the machine reference only in how the words *committee* and *appointment* are translated. The human word choice is fluent but has a lower probability under the model.

The second example shows that machine references often avoid the phenomenon altogether. Here, a simple conjunction is used instead of the more prestigious preposition *mitsamt* 'along with' in the human reference. This removes any risk of inserting a hypercorrect genitive. Since a contrastive variant cannot be derived from the machine reference, the sample is excluded from the machine-generated test set.

C   State-of-the-art Accuracies for DistilLingEval

Error Type	Test set type	Discrepancy of test set	Reported accuracy
clause_omission	human references	0.91	78.1
	machine references	0.19	87.1
hypercorrect_genitive	human references	1.13	94.2
	machine references	0.18	100.0
np_agreement	human references	0.84	99.7
	machine references	0.15	100.0
placeholder_ding	human references	0.79	99.8
	machine references	0.16	100.0
polarity_affix_del	human references	0.87	98.8
	machine references	0.13	100.0
polarity_particle_kein_del	human references	0.88	97.5
	machine references	0.12	100.0
polarity_particle_nicht_del	human references	0.84	97.9
	machine references	0.14	99.9
subj_verb_agreement	human references	0.83	98.7
	machine references	0.14	99.8

Table 6: DistilLingEval results of a state-of-the-art NMT ensemble (Ng et al., 2019). The accuracies on machine references suggest that clause omission is an error type that still occurs with state-of-the-art NMT systems.