Contrastive Conditioning for Assessing Disambiguation in MT: A Case Study of Distilled Bias

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

Lexical disambiguation is a major challenge for machine translation systems, especially if some senses of a word are trained less often than others. Identifying patterns of overgeneralization requires evaluation methods that are both reliable and scalable. We propose *contrastive conditioning* as a reference-free black-box method for detecting disambiguation errors. Specifically, we score the quality of a translation by conditioning on variants of the source that provide contrastive disambiguation cues. After validating our method, we apply it in a case study to perform a targeted evaluation of sequence-level knowledge distillation. By probing word sense disambiguation and translation of gendered occupation names, we show that distillation-trained models tend to overgeneralize more than other models with a comparable BLEU score. Contrastive conditioning thus highlights a side effect of distillation that is not fully captured by standard evaluation metrics. Code and data to reproduce our findings are publicly available.[1]

1 Introduction

Erroneous disambiguation of words makes translations inadequate and can even constitute a form of bias when it occurs systematically. However, detecting disambiguation errors in machine translation (MT) is a non-trivial task. Previous work has focused on automatic post-hoc analysis of translations (Raganato et al., 2019; Stanovsky et al., 2019), but rules of what makes a disambiguation correct or incorrect tend to be imprecise. While *contrastive evaluation* (Sennrich, 2017; Rios et al., 2017) eliminates the need for post-hoc analysis by scoring pre-defined pairs of hypotheses, such probability estimates cannot be obtained from black-box systems, e.g., from commercial APIs that only return a 1-best translation to the user. In addition, contrastive hypotheses need to be carefully crafted for every target language of interest.

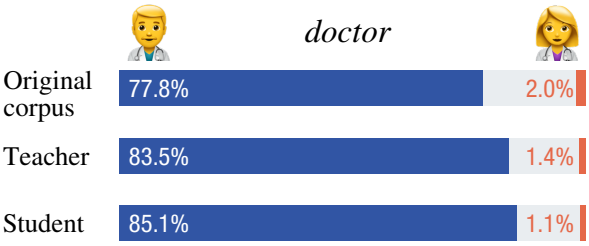

Figure 1: Our case study is motivated by an analysis of the training data: In the English–German WMT19 news corpus, *doctor* is mostly translated into male forms such as *Arzt* and rarely into female forms (center: other variants). Student models trained on a machine-translated version of the data amplify this imbalance.

We propose *contrastive conditioning* as a scalable black-box alternative for evaluating disambiguation in machine translation. The evaluated translations are paired with contrastive source sequences and are then scored by a white-box translation model. The contrastive sources are variants of the original source, slightly modified to provide a stronger disambiguation cue. For example, consider a model that translates the English source *'doctor'* into German as *'Ärztin'* (female doctor). This translation will receive a better score when conditioned on the source *'female doctor'* than on *'male doctor'*, indicating that it is a feminine form. Given sufficient disambiguation cues, the white-box translation model can thus serve as an evaluator for the original translation.

Since the contrastive sources are written in the source language, contrastive conditioning does not rely on references in the target language. This makes it easier to scale the evaluation across multiple target languages. In addition, the method is reliable compared to alternative evaluation methods, according to our human validation.

In a case study, we utilize contrastive condition-

[1]https://github.com/ZurichNLP/contrastive-conditioning

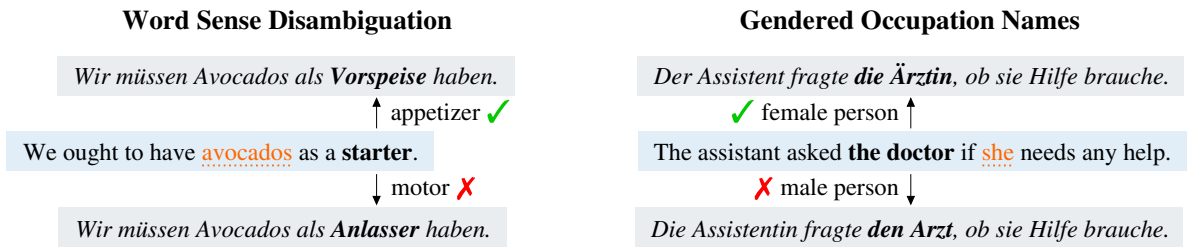

Figure 2: Examples of the MuCoW (left; Raganato et al., 2019) and WinoMT (right; Stanovsky et al., 2019) tasks in contrastive format. The German translations illustrate a disambiguation problem: Only those on the top correctly disambiguate the ambiguous source word (printed in bold) given the relevant context (underlined in orange).

ing to answer a specific research question. We probe models trained using sequence-level knowledge distillation (SeqKD; Kim and Rush, 2016) and quantify their *overgeneralization bias*, i.e., their tendency to err on the side of frequent word senses. This question is of emerging interest because the distilled training data used for SeqKD are known to have reduced entropy (Zhou et al., 2020). Figure 1 illustrates to what degree a rare word sense can vanish in distillation, raising the question of how this affects disambiguation quality.

Our case study is based on English–German and English–Russian systems and applies contrastive conditioning to two distinct types of disambiguation. The first type is word sense disambiguation in general, as represented by the MuCoW test suite (Raganato et al., 2019). The second type is the special case of gendered occupation names, for which the WinoMT challenge set has been released (Stanovsky et al., 2019). For both types of disambiguation, our results based on contrastive conditioning confirm that models trained via SeqKD tend to have a more pronounced overgeneralization bias than other models with a comparable BLEU score.

2 Background and Related Work

2.1 Evaluation of Disambiguation in MT

In the context of translation, word sense disambiguation (WSD; Navigli, 2009) can be formally defined as follows: Let us assume that every instance of a content word w conveys one out of a set $\{s_1, s_2, \dots\}$ of senses. Then a WSD error occurs if a source instance w_i is translated into a target word that does not convey the sense of w_i but another sense of the word w (Popescu-Belis, 2019).

Automated approaches for evaluating MT systems on WSD can be grouped into *pattern-matching* and *scoring* approaches. In a pattern-matching evaluation, translations are searched for phrases that are known to be correct or incorrect. For example, Vickrey et al. (2005) create a test set from ambiguous source words in a parallel corpus, and Raganato et al. (2019, 2020) use this approach to assemble a large-scale benchmark (MuCoW) with multiple translation variants for ambiguous words. However, it is usually not feasible to create an exhaustive list of all translation variants.

Scoring-based evaluation alleviates this problem by directly comparing probabilities for pre-defined contrastive translation variants as estimated by the model (Sennrich, 2017; Rios et al., 2017). An example of a contrastive translation pair for WSD is presented in the left part of Figure 2. The scoring of contrastive translations has some drawbacks in that it depends on a non-standard interface to the MT system, and, like pattern-matching evaluation, on language-specific references. Furthermore, there is no guarantee that the actual 1-best translation would be similar to the preferred variant.

2.2 Translation of Gendered Occupations

The translation of gendered occupation nouns can be seen as a special case of WSD. When translating occupations from a language that does not tend to mark their gender into a language that does, gender has to be either inferred from the context, e.g. from anaphoric pronouns, or expressed neutrally. Such a challenge arises when translating from English into German, Russian or other morphologically rich languages. WinoMT (Stanovsky et al., 2019) is a challenge set for this phenomenon, which combines several datasets for gender coreference in English (Rudinger et al., 2018; Zhao et al., 2018). See Figure 2 for an example in contrastive format.

2.3 Overgeneralization Bias

Carbonell et al. (1983) describe *overgeneralization* as a tendency to learn concepts that extend not only to positive but also to negative examples, which

can arise if a system sees mostly positive examples. More recently, overgeneralization has been discussed as a category of social impact of NLP systems (Hovy and Spruit, 2016), and it has been hypothesized that overgeneralization of the training data leads to a loss of lexical diversity and to an exacerbation of gender bias in MT (Vanmassenhove et al., 2019; Roberts et al., 2020). In the case of WSD, Rios et al. (2017) have found that neural MT systems handle frequent word senses well but perform poorly on rare word senses. The influence of sense distribution on WSD has been further examined by Tang et al. (2018), Raganato et al. (2020) and Emelin et al. (2020). With regard to gendered occupations, Stanovsky et al. (2019) show that MT translates stereotypes more reliably, and WinoMT or similar datasets have subsequently been used to quantify bias in various translation settings (Kocmi et al. 2020; Costa-jussà et al. 2020a,b; Tomalin et al. 2021; Choubey et al. 2021; Renduchintala and Williams 2021; among others).

2.4 Effects of Knowledge Distillation

Overgeneralization bears some resemblance to compression, which is significant in the context of knowledge distillation (Hinton et al., 2015). The process of sequence-level knowledge distillation (SeqKD) can be described as follows (Kim and Rush, 2016):

1. A generative model is trained, to be used as an intermediate *teacher*;
2. The teacher re-generates the target side of the training data;
3. A *student* model, which is usually smaller, is trained on the new data.

In its simplest form, SeqKD replaces the original target side of the training data with the teacher's best translation as generated with beam search. Kim and Rush (2016) report that small student models can approximate the translation quality of more complex teachers and that student models excel under greedy decoding, making them an attractive choice for large-scale deployment of MT (Kim et al., 2019). The effectiveness of SeqKD raises the question of how distilled data differ from the original training data and how such a difference might affect model behavior.

Previous analyses of SeqKD have focused on general linguistic metrics rather than probing tasks such as lexical disambiguation. Distilled data have been characterized as less noisy and more deterministic than the original target (Gu et al., 2018), as having a reduced fertility and distortion (Zhang et al., 2018), reduced lexical diversity (Xu et al., 2021), and as being less complex while preserving faithfulness (Zhou et al., 2020).

Concurrent work (Silva et al., 2021) studies distillation in the context of masked language models, showing that distilled models have a more pronounced bias according to standard metrics. Ding et al. (2021) examine SeqKD in non-autoregressive MT, where it is shown to decrease translation accuracy with respect to rare words in word-aligned parallel data. Finally, Renduchintala et al. (2021) show that MT models optimized for speed have an increased gender bias, analyzing techniques that are complementary to distillation, namely reduction of beam size, shallow decoders, efficient attention and quantization. In this paper, we perform a case study on distillation based on established probing tasks in MT, using a novel evaluation protocol in order to reliably identify patterns of overgeneralization.

3 Contrastive Conditioning

3.1 Linguistic Motivation

Recall of Pattern-matching Approaches Evaluation methods for disambiguation can have limited recall, which adds noise to comparisons of related systems. For example, Scherrer et al. (2020) find that on average, 20–28% of translations remain undecidable given the MuCoW gold variants.

WinoMT follows a pattern-matching approach as well, but without using reference translations. The predicted gender is inferred based on word alignment and language-specific morphosyntactic analysis. Stanovsky et al. (2019) have found that such an analysis yields an average agreement with human annotations of 87%, and manual whitelisting was proposed to further improve accuracy (Kocmi et al., 2020). However, a certain amount of noise cannot be avoided. In a small percentage of cases, the analyzers cannot determine the grammatical gender; Kocmi et al. (2020) counted around 13% such unresolved translations for Russian. In addition, errors in alignment or morphosyntactic analysis can cause a small number of false positives or negatives.

Classification of Gender In some cases, precision of WinoMT is impacted by the *equation of notional and grammatical gender* when an English

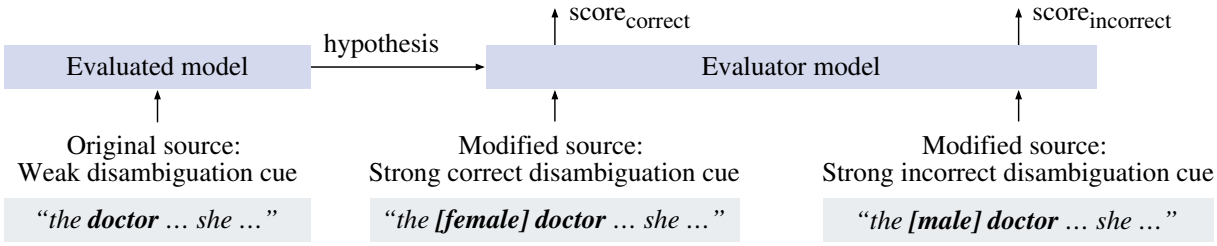

Figure 3: Schema of *contrastive conditioning*, which can be used to evaluate disambiguation in MT.

pronoun is compared to the morphosyntactic properties of a translated noun. However, an English pronoun conveys the notion of a speaker about a real-world referent (*notional gender*; McConnell-Ginet, 2014), while in languages such as German, Romance languages or Russian, the gender of nouns is sometimes arbitrary (*grammatical gender*; Bender, 2013). For example, the German noun *Wache* 'guard' is grammatically feminine but can refer to any person. Thus, morphosyntactic analysis, despite being a good heuristic in general, can lead to classification errors regarding gender.

Furthermore, a disagreement between notional and grammatical gender can at times be interpreted as a generic, rather than false, translation. However, this seems unlikely for WinoMT because most sentences describe concrete individuals of known gender. A notable exception may be Russian, where masculine nouns are used for many occupations – e.g., врач 'doctor' – and choice of grammatical gender can be influenced by factors such as prestige or historical connotation in addition to the referent's gender (Wade, 2011). Finally, an exclusive focus on grammatical gender may penalize efforts to create gender-neutral translations. For example, neutral terms in German tend to coincide with a feminine grammatical gender, such as *Pflegekraft* 'care worker' or *Fachperson* 'specialist person'.

3.2 Proposed Evaluation Protocol

Given those considerations, we propose an alternative evaluation protocol that does not impose hard constraints on a translation (Figure 3):

1. Translate the original source with the model that is being evaluated.
2. Construct variants of the source that provide a stronger disambiguation cue. The variants are contrastive: Some disambiguate the source correctly, others do so incorrectly.
3. Use a translation model (*evaluator*) to score the translation from (1) conditioned on the contrastive sources. Compute $\text{score}_{\text{correct}}$ as the best evaluator score over the correctly modified sources, and $\text{score}_{\text{incorrect}}$ as the best evaluator score over the incorrectly modified sources.[2]

The overall score for the translation is defined as:

$$\text{score} = \text{score}_{\text{correct}} / (\text{score}_{\text{correct}} + \text{score}_{\text{incorrect}}).$$

Note that the evaluated model could be used as its own evaluator. To make comparisons between many models, however, we prefer to evaluate all models with the same state-of-the-art ensemble.

3.3 Disambiguation Cues

The disambiguation cues used for contrastive conditioning can have textual form, or have the more generic form of an additional input feature (Sennrich and Haddow, 2016). In our experiments, we insert a textual cue into the source sentence because this enables us to use an off-the-shelf MT model as evaluator.

MuCoW We prepend the ambiguous word with another word that hints at a specific sense:

> Original source: *We ought to have avocados as a **starter**.*
>
> Modified source (correct): *We ought to have avocados as a **dinner starter**.*
>
> Modified source (incorrect): *We ought to have avocados as a **motor starter**.*

In order to automatically find disambiguation cues to insert, we use a masked language model (Liu et al., 2019) to generate a set of candidates for each sense. In a second step, we select the insertions that prove to be most discriminative for contrastive conditioning; we use the reference translations provided by MuCoW as a validation set. We select 3 correct and 3 incorrect insertions per sense. Implementation details are provided in Appendix D.

[2]The definition of a score can vary between MT systems. In this paper, we use the average positional likelihood across all tokens of the hypothesis, so higher scores are better.

No.	Translation	Gold rating	score correct	score incorrect	score
1	*Der Assistent fragte **die Ärztin**, ob sie Hilfe brauche.*	✓ correct	.794	.745	.516
2	*Der Assistent fragte **die Doktorin**, ob sie Hilfe brauche.*	✓ correct	.645	.599	.519
3	*Die Assistentin fragte **den Arzt**, ob sie Hilfe brauche.*	✗ incorrect	.765	.814	.484
4	*Die Assistentin fragte **den Doktor**, ob sie Hilfe brauche.*	✗ incorrect	.668	.710	.485
5	*Die Assistenz fragte **die ärztliche Fachperson**, ob sie Hilfe brauche.*	ambiguous	.404	.393	.507
6	*Die Assistentin fragte, ob sie Hilfe brauche.*	ambiguous	.610	.607	.501

Table 1: Some potential German translations of *"The assistant asked the doctor if she needs any help"* and their contrastive conditioning scores. The scores have been computed by an English–German ensemble (Ng et al., 2019) based on contrastive sources. The example illustrates that contrastive conditioning is informative: The first two translations (#1–2), which disambiguate *doctor* correctly while differing in word choice, receive an overall score > 0.5. On the other hand, the two incorrect translations (#3–4) receive a score < 0.5, and the overall score is close to 0.5 for translations that are ambiguous either due to gender-neutral language (#5) or word omission (#6).

Gendered Occupations We add the adjectives 'female' and 'male' in brackets:

> Original source: *The assistant asked the **doctor** if she needs any help.*
>
> Modified source (correct): *The assistant asked the **[female] doctor** if she needs any help.*
>
> Modified source (incorrect): *The assistant asked the **[male] doctor** if she needs any help.*

We treat the disambiguation cue that agrees with the WinoMT gold label as correct, and vice versa.

3.4 Weighting of Samples

Unweighted Accuracy In the simplest form, we can define the accuracy of the evaluated model to be the proportion of samples where $\text{score} > 0.5$.

Category-wise Weighted Accuracy However, the evaluator score can be interpreted as a form of *confidence*, since the likelihood that contrastive conditioning misjudges a translation is highest where $\text{score} \approx 0.5$. We propose to downweight samples where the evaluator has low confidence, using a weighting scheme that maintains the balance of categories. For each category (e.g. 'male'), we rank the samples in decreasing order by $|\text{score} - 0.5|$, and assign to each sample i a weight proportional to $n - rank(i)$, where n is the size of the category. We use the weights to compute a weighted accuracy.

Table 1 illustrates the scoring process on the example of hand-crafted translations, and in Tables A5 and A6, further examples of real machine translations are discussed. We find that translations with low evaluator confidence tend to be difficult to judge, either because they carry over the ambiguity, or because the ambiguous part of the source has been omitted in the translation.

Method	MuCoW DE	MuCoW RU	WinoMT DE	WinoMT RU
Pattern-matching baseline	83.8	89.7	86.3	75.5
Contrastive conditioning	80.6	85.8	92.7	72.7
– weighted	88.0	92.8	97.6	78.2

Table 2: In a human validation study, we compare different automated evaluation methods to human evaluation of machine translations. Each figure is the *proportion of agreement* between human evaluation and automated evaluation. The agreement in the last row is weighted according to category-wise weighting.

3.5 Validation of Contrastive Conditioning

Human Validation We perform a human validation study to verify that contrastive conditioning is a viable alternative to the pattern-matching baselines. English–German and English–Russian machine translations, together with the sources, are blindly annotated by 2 language professionals (up to 200 samples per language, task and annotator). For MuCoW samples, we ask: *Is the translation closer to the correct sense cluster than to the wrong one?* For WinoMT, the question asked is: *Does the occupation name convey the gender implied by its context?*

The inter-annotator agreement is in the range expected for the tasks, ranging from $\kappa = 0.95$

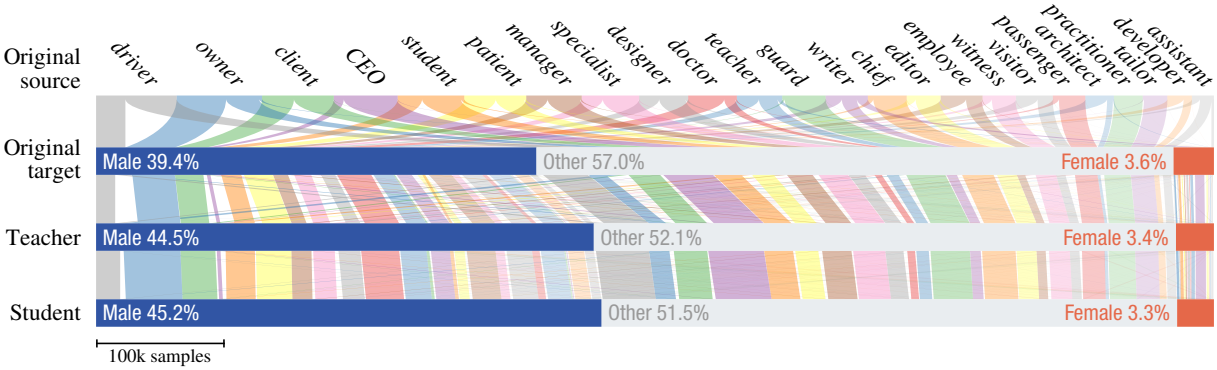

Figure 4: A Sankey diagram visualizing how sequence-level knowledge distillation amplifies gender imbalances for the most frequent occupations in the WMT19 English–German corpus. Redistribution of grammatical gender is observed mostly when the teacher regenerates the target side of the training data, but the student further increases the imbalance when prompted to translate the same data. Unidentified translation variants are classified as 'Other'.

for English–German WinoMT to $\kappa = 0.20$ for English–Russian WinoMT. The low agreement for the latter task confirms that the degree to which Russian occupation nouns are generic is highly subjective, as discussed in Section 3.1.

Our results (Table 2) show that contrastive conditioning has a high proportion of agreement to the human annotations. Details of the human validation study are described in Appendix E.

Agreement of Different Evaluator Models As an additional validation of our method, we analyze if the choice of evaluator model can influence the results. Namely, we rank all the 27 models evaluated in the case study (Section 4) with 3–4 state-of-the-art-models that have been independently trained with various random seeds (Ng et al., 2019). The evaluators largely agree in their rankings. When averaging the Spearman's rank correlation coefficients over all pairs of evaluators for each task and language pair, the maximum average is 0.99 for English–German WinoMT, and the minimum is 0.88 for English–Russian WinoMT. In the latter case, a lower rank correlation is expected given that some of the evaluated models perform very closely.

3.6 Effect on Gender-Neutral Language

The WinoMT dataset includes a small number of source sentences that contain the pronoun *they*. Such examples show that translation of gender cannot solely be understood as a disambiguation problem, and has more complex aspects. In this study, we follow Kocmi et al. (2020) and exclude neutral inputs from the dataset since we use the dataset as a proxy to quantify disambiguation error. However, contrastive conditioning does not depend on binary labels, and we hope that the proposed method could also aid the assessment of gender-neutral or non-binary translation (Cao and Daumé III, 2020; Saunders et al., 2020). Furthermore, it was mentioned in Section 3.1 that gender-neutral language can be a valid way to translate ambiguous source sentences, but one that, in some languages, is difficult to evaluate based on grammatical gender. While our approach does not directly recognize neutral language, such edge cases can likely be identified by a low evaluator confidence. We downweight cases with low confidence using the above weighting scheme, given that within the framework of WSD, translations that preserve the ambiguity of the source are usually not considered to be disambiguation errors (Popescu-Belis, 2019).

4 Case Study: Assessing Disambiguation Bias in Distilled Models

4.1 Hypothesis

We hypothesize that distillation could also impair disambiguation quality. Our hypothesis is motivated by a simple data analysis, which is visualized in Figure 4.

Motivating Analysis When searching for English occupation nouns in the English–German parallel training data from the WMT19 news translation task (Barrault et al., 2019), we find that the gender ratio is considerably skewed: Of the corresponding German references, 39.4% contain common translation variants that convey the notion of a male person (e.g., *Fahrer* 'male driver') and only 3.6% contain common female variants (e.g., *Fahrerin* 'female driver'); the remaining 57.0%

contain neutral, impersonal, or lexically rare translations. In addition to real-world labor inequalities, this skew can be explained by linguistic phenomena such as *generic masculines* (Lessinger, 2020).

As shown in Figure 4, we also analyze the translations across two additional iterations of the data: the distilled data generated by the teacher, and translations of the same training data by the student model. We average our counts across three teachers and three students trained with different random seeds, which allows us to report standard deviations. As expected, the distilled data have a higher imbalance, with male forms increasing by 5.1% ($\pm 0.4\%$) and female forms decreasing by 0.2% ($\pm 0.0\%$). Moreover, the student further increases this imbalance (despite having the same size and capacity as the teacher), with male forms growing by 0.7% ($\pm 0.2\%$) and female forms slightly decreasing by 0.1% ($\pm 0.1\%$). It seems plausible that smaller students would develop an even stronger bias.

Limitations of Data Analysis However, such a word count analysis of distillation has clear limitations (we describe implementation details in Appendix A and address further limitations in the impact statement). One limitation is that there is a large number of translations that cannot be automatically classified. Focusing on precision not only leaves a very large group of translations classified as 'Other' but could itself be a source of bias.

Secondly, the word count analysis does not take into account whether a source sentence provides sufficient context for disambiguation, or whether a source would be inherently ambiguous even to human translators. While it seems likely that the lexical overgeneralization observed in the above analysis will also cause translation errors, such errors can only be quantified using source sentences that are known to have a salient context. Our preliminary data analysis as well as its potential limitations thus strongly motivate a targeted evaluation of SeqKD models with respect to disambiguation.

4.2 Experimental Setup

To have a controlled setup, we trained teachers and students from scratch using Fairseq (Ott et al., 2019). In addition, we also distilled state-of-the-art MT systems (Ng et al., 2019).[3] The main difference between the 'Scratch' and the 'SOTA' teachers is that Ng et al. (2019) used advanced filtering, backtranslation and domain-adaptive fine-tuning.[4]

Architecture For the teachers, we used the big Transformer architecture (Vaswani et al., 2017) with a doubled feed-forward size of 8192. For the students and for further baselines we trained two additional sizes, *small* and *mini*. Table A1 compares the three sizes and their parameter numbers.

Data To train our teachers we used the English–German and English–Russian parallel training data from the WMT19 news translation task (Barrault et al., 2019).[5] We reused the BPE vocabularies computed by Ng et al. (2019), which are joint for English–German and disjoint for English–Russian. We also filtered sentences longer than 250 tokens as well as pairs with a length ratio larger than 1.5.

Hyperparameters For each language pair we trained 3 'Scratch' teachers with different random seeds. Each teacher was then used to train an individual student per size. We repeated this procedure with the 'SOTA' teachers. We used Adam with an initial learning rate of 5e-4, FP16 training, label smoothing with $\epsilon = 0.1$, and a dropout of 0.3. We trained with a token batch size of 16k, and we selected the best checkpoint with respect to BLEU based on the *newstest* sets from the preceding years.

For decoding, beam search with size 5 was used.

Evaluation To evaluate general translation performance we used the WMT19 testset and computed BLEU (Papineni et al., 2002) using SacreBLEU (Post, 2018).[6] To evaluate disambiguation accuracy we used the MuCoW and WinoMT test sets. In the case of MuCoW, the WMT19 version of the complete translation-based data was used (Raganato et al., 2019). Details of those datasets are reported in Table A2. Finally, we used the ensembles by Ng et al. (2019) as an evaluator for contrastive conditioning and applied category-wise weighting (Section 3.4).

Metrics Since our goal is to quantify overgeneralization bias, a metric is required that captures overgeneralization based on disambiguation test-

[3] We considered four individual teachers per language pair that have been released as an ensemble by Ng et al. (2019). However, the first English-German submodel of the ensemble is identical to the third one, so we only used three as teachers.

[4] The WMT19 submission of Ng et al. (2019) also involves reranking, which we do not use in our experiments.

[5] We used ParaCrawl Corpus release v5.0 instead of v3.0.

[6] The version was BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+test.wmt19+tok.13a+version.1.4.14.

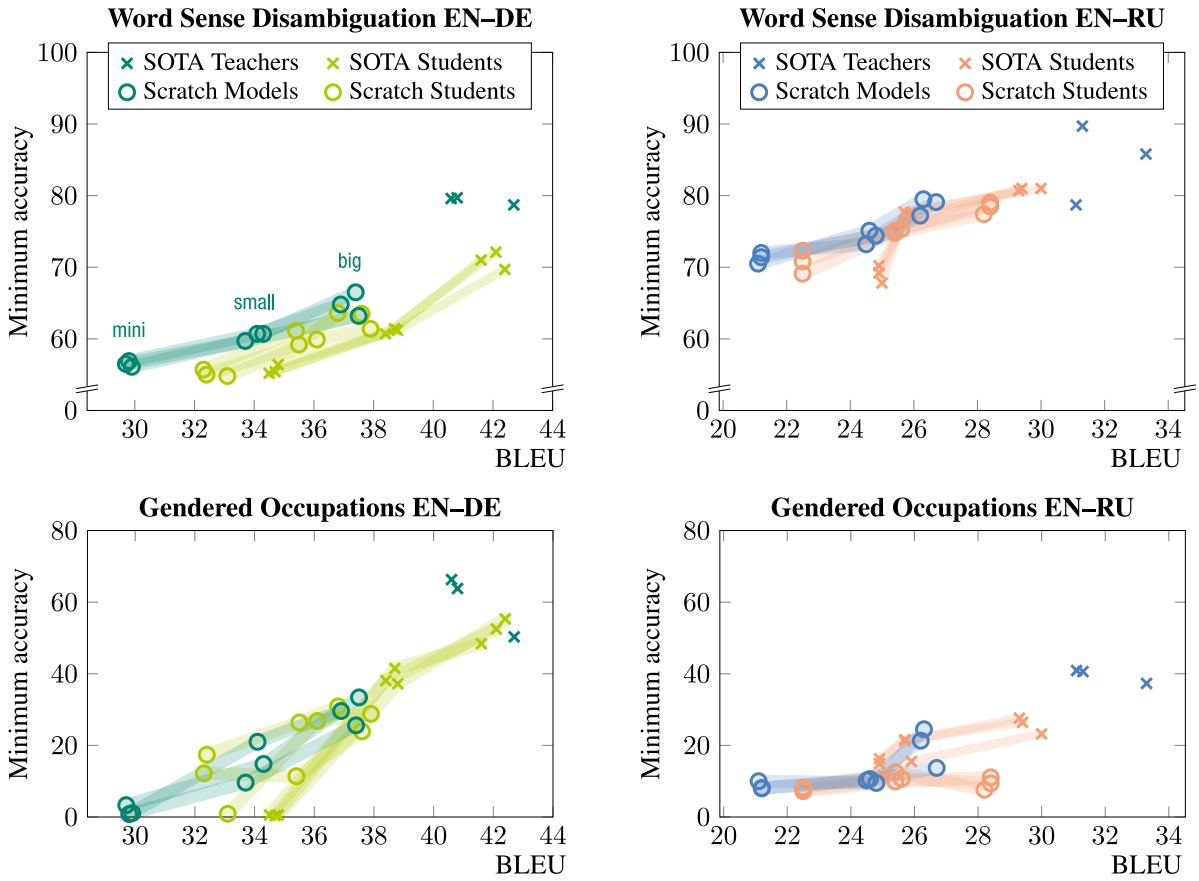

Figure 5: Performance of English–German (left) and English–Russian (right) translation models in terms of BLEU on *newstest19*, and MuCoW (top) as well as WinoMT (bottom) minimum accuracy. In order to highlight compression effects, students of varying capacity that share the same teacher are connected with a line. Additional lines connect baseline models of varying capacity that were trained from scratch with the same random seed.

sets such a MuCoW and WinoMT. We use *minimum accuracy* over the categories of those testsets.

For MuCoW, we use the minimum accuracy over the categories 'frequent' and 'rare'. Word senses are categorized as 'frequent' if they occur more often than alternative senses in the training data. The other, alternative senses are categorized as 'rare' (our word counting is described in Appendix A).

For WinoMT, we use the minimum accuracy over the categories 'male' and 'female'. This is a deviation from what has been used in previous work on WinoMT, but we believe that it serves as an adequate measure of overgeneralization. While in theory, minimum accuracy is motivated by the difference principle of distributive justice (Rawls, 1971), in practice we find that minimum accuracy is consistently found in the categories 'rare' (MuCoW) and 'female' (WinoMT). This confirms that minimum accuracy captures overgeneralization bias. Absolute differences (Δ_G or Δ_S; Stanovsky et al., 2019) or a ratio of categories (M:F; Saunders and Byrne, 2020) do not take overall performance into account and thus assign good scores to models with low accuracy. In addition, minimum accuracy is easy to interpret, ranging from 0 to 100, with higher meaning better.

4.3 Results

Figure 5 shows our results, which are listed in tabular form in Appendix H. While BLEU is positively correlated with minimum accuracy according to our overgeneralization probing tasks, student models tend to perform worse on the probing tasks than other models with a similar BLEU score. In order to statistically confirm this observation, we perform a multiple linear regression analysis for each task–language pair (Table 3). We find that BLEU has a significant positive correlation to accuracy on the overgeneralization probing tasks and that SeqKD has a significant negative correlation. The overall regression is statistically significant ($p < 0.05$).

We also note that for all English–Russian mod-

Task	Variable	Coefficient
MuCoW EN–DE	BLEU score	1.78*
	SeqKD is used	-6.47*
MuCoW EN–RU	BLEU score	1.39*
	SeqKD is used	-2.14*
WinoMT EN–DE	BLEU score	5.00*
	SeqKD is used	-7.56*
WinoMT EN–RU	BLEU score	2.52*
	SeqKD is used	-5.21*

Table 3: A multiple regression analysis confirms that students trained with SeqKD tend to perform worse on probing tasks for overgeneralization. The dependent variable is *minimum accuracy* on the probing task; as independent variables the BLEU score is used, as well as a binary variable describing whether the model was trained as a SeqKD student (*: significant at $p < 0.05$).

els, the minimum accuracy on gendered occupations is worse than random, which is in line with previous findings by Kocmi et al. (2020). Remarkably, English–Russian teachers trained from scratch are far outperformed by their best students in terms of BLEU, but the students are more biased towards overgeneralization.

5 Discussion

Contrastive conditioning is a new protocol for evaluating and comparing MT systems with regard to word sense disambiguation. In our analysis we build on established test sets, but have replaced the standard pattern-matching analysis with contrastive conditioning. A human validation study, with English as a source language, showed that the approach is reliable, especially if the samples are weighted according to evaluator confidence.

An advantage of our approach over pattern-matching is that it can process any potential translation, not just translations containing pre-defined lemmas or translations that have certain morphosyntactic properties. Furthermore, an advantage over conventional contrastive evaluation methods is that the decoding mode of the evaluated model does not need to be constrained. Thus, black-box systems, for example APIs of commercial systems, can be evaluated too. Finally, test sets can remain reference-free, and we even believe that neither strong assumptions nor deeper expertise of a target language are strictly required to perform contrastive conditioning (even though in this paper, we put forward some linguistically informed arguments to motivate our approach).

A limitation of contrastive conditioning is that a disambiguation cue needs to be provided in the source language. For error types other than disambiguation, such a cue might be difficult to create. In this paper, we have built on purely textual cues, which enabled us to use an off-the-shelf translation model for scoring. Linguistic input features (Sennrich and Haddow, 2016; Stafanovičs et al., 2020) could provide an alternative disambiguation cue in future work.

Our case study of knowledge distillation, which is based on contrastive conditioning, shows that SeqKD can lead to lexical overgeneralization, and to a loss of adequacy in disambiguation that is generally not captured by BLEU.

6 Conclusion

In order to evaluate MT models on disambiguation, we have devised a novel evaluation method, *contrastive conditioning*. It allows for a reference-free, black-box evaluation of MT models with respect to disambiguation, requiring only that a strong disambiguation cue can be provided in contrastive sources. Based on this evaluation method, we have presented a case study of translation models trained with sequence-level knowledge distillation. Focusing on the issue of lexical overgeneralization in word sense disambiguation, we have tested the models on word sense disambiguation and the translation of gendered occupations. Our results indicate that sequence-level knowledge distillation can amplify existing imbalances in the training data, and typically leads to an increased overgeneralization bias. We encourage future work to develop methods that reap the benefits of knowledge distillation with minimal increases in bias.

Acknowledgments

This work was funded by the Swiss National Science Foundation (project MUTAMUR; no. 176727) and made use of infrastructure services provided by S3IT, the Service and Support for Science IT team at the University of Zurich. We would like to thank Denis Emelin, Anastassia Shaitarova and Wanda Siegenthaler for providing comments and annotations, and Annette Rios, Gabriel Stanovsky as well as the anonymous reviewers for helpful feedback. We also thank the ACL Rolling Review team for their effort in organizing a pilot run.

Broader Impact

We use the term 'bias' to describe a behavioral tendency of NLP systems that goes undetected by common evaluation metrics. While we focus on how it affects the accuracy of translations, overgeneralization bias does not just have a technical dimension but also a social one (Hovy and Spruit, 2016; Sheng et al., 2021), especially with regard to sensitive categories such as gender. Therefore, our findings could also inform a socio-political discussion of model compression, provided that such a discussion is normatively well-founded (Blodgett et al., 2020).

Our preliminary data analysis (Section 4.1) is based on gender as a variable, which warrants some ethical reflection (Larson, 2017). Our analysis is based on a very large collection of English occupation nouns and their translations into German. We categorize the notional gender of the German translations as 'male' or 'female' in cases where grammatical gender is a valid indication. While the automatic inference of gender is discouraged in many research contexts, we believe that our approach is adequate, since in this case, rather than the personal gender of human subjects, the notional gender of nouns is inferred (McConnell-Ginet, 2014). However, gender-neutral or alternative ways of expressing gender are not separately counted. Thus, the preliminary data analysis should be understood as a motivating example of lexical overgeneralization, and does not constitute a comprehensive corpus analysis of gender.

References

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.

Emily M Bender. 2013. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies*, 6(3):1–184.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Jaime G. Carbonell, Ryszard S. Michalski, and Tom M. Mitchell. 1983. An overview of machine learning. In Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell, editors, *Machine Learning. An Artificial Intelligence Approach*, pages 3 – 23. Morgan Kaufmann, San Francisco (CA).

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. Improving gender translation accuracy with filtered self-training. *arXiv preprint arXiv:2104.07695*.

Marta R Costa-jussà, Christine Basta, and Gerard I Gállego. 2020a. Evaluating gender bias in speech translation. *arXiv preprint arXiv:2010.14465*.

Marta R Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020b. Gender bias in multilingual neural machine translation: The architecture matters. *arXiv preprint arXiv:2012.13176*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *International Conference on Learning Representations*.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. 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.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Enora Lessinger. 2020. The challenges of translating gender in UN texts. *The Routledge Handbook of Translation, Feminism and Gender*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sally McConnell-Ginet. 2014. Gender and its relation to sex: The myth of 'natural' gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

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.

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.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. *arXiv preprint arXiv:1901.09115*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3668–3675, Marseille, France. European Language Resources Association.

John Rawls. 1971. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Massachussets.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural 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 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2021. Investigating failures of automatic translation in the case of unambiguous gender. *arXiv preprint arXiv:2104.07838*.

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.

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. 2020. Decoding and diversity in machine translation. In *Resistance AI Workshop*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The MUCOW word sense disambiguation test suite at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 365–370, Online. 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.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. 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.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Artūrs Stafanovičs, Mārcis Pinnis, and Toms Bergmanis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics.

Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. 2021. The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics and Information Technology*.

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.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 771–778, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Terence Wade. 2011. *A comprehensive Russian grammar*. John Wiley & Sons.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4392–4400, 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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

A Word Count Methodology

There are two instances where we count word occurrences in the training data:

- We count the senses of ambiguous English words in order to divide the samples into the categories 'frequent' and 'rare'. For this we use an approximative method.
- We count the genders of German occupation names to inform Section 4.1 and Appendix C. We make sure to use a method with high precision for this.

Word Senses For an approximate count of English word senses we use a similar method as Raganato et al. (2020). The MuCoW dataset represents a sense of an English source word as a cluster of target-language lemmas. We thus count a sentence pair as an occurrence of a sense if the source word appears in the source and at least one of the target lemmas appears in the lemmatized reference. We lemmatize the data using Stanza (Qi et al., 2020). The count is approximate since (a) the provided variants in the target language do not cover all possible translations and (b) the lemmatization is noisy. Still, we expect the counts to be proportional to the true sense distribution.

Occupation Names To count the genders of German occupations, we list common German translations of each English occupation name. We only list variants that have an identifiable gender across the full morphological paradigm, and whose grammatical gender usually matches the notional gender. (Most German occupational terms meet this criterion, but there are exceptions such as *Angestellte*, whose male and female paradigms intersect, and the gender-neutral *Wache* mentioned above.) On average, we list 3–4 male lemmas per occupation, and the same amount of female variants. For each lemma, we enumerate the complete paradigm and search the data for each inflectional form. Note that masculine occupation nouns usually have more inflectional forms than feminine nouns, but we do not expect this to influence our results since the totals over the full paradigm should be comparable. We count each sentence pair as an occurrence of 'male' if the English occupation is found in the source and one of the male forms is found in the target sequence. If one of the female forms is found, we count the occurrence as 'female', and if no known forms, or both male and female forms, are found we classify the translation as 'other'.

B Model Sizes

Name	N	d_{model}	d_{ffn}	h	Parameters
big	6	1024	8192	16	269.7M
small	4	512	2048	4	50.9M
mini	4	256	1024	4	18.1M

Table A1: Transformer sizes used for students and non-distilled baselines.

C Occupational Stereotypes

Since WinoMT also uses a notion of stereotypes (Δ_S), we considered using this metric for our analysis. However, the top 25% of occupations in the English–German training data are all predominantly associated with male word forms. In the top 50% occupations, which together have a relative frequency of 95%, there are just two occupations that are mostly associated with female forms in the data (*nurse* and *cleaner*), We did find some correlation between the female ratios in the training data and the percentages derived by Zhao et al. (2018) from U.S. labor statistics, with a Pearson coefficient of $r = 0.69$. Still, since the predominant stereotype in the German training data is 'male' for all but 2 occupations that we searched, we did not extend our analysis to occupational stereotypes.

D Disambiguation Cues for WSD

After some experimentation, we decided to rely on a fully automated approach for finding suitable insertions, which involves a masked language model to generate inserted words, and a validation process to select the most discriminative insertions. Insertions are generated based on the MuCoW source sentences. For every sentence, we place a `<mask>` token before the ambiguous word and predict a probability distribution using RoBERTa (Liu et al., 2019). For each sense cluster, we select the 10 words with the highest predicted probability of occurring in the example sentences but not in the counterexamples, and 10 words vice versa. We then use the reference translations provided by MuCoW as a validation set to reduce these potential disambiguation cues to the 3 correct and 3 incorrect cues that are most discriminative for contrastive conditioning. Correct disambiguation cues are discriminative if the evaluator assigns a good score to the reference, and vice versa.

Dataset	URL	Number of Samples
WMT19 News Translation Task EN–DE (with ParaCrawl v5.0 instead of v3.0)	http://www.statmt.org/wmt19/translation-task.html	Train: 44 349 092 Validation: 30 746 Test: 1997
WMT19 News Translation Task EN–RU (with ParaCrawl v5.0 instead of v3.0)	http://www.statmt.org/wmt19/translation-task.html	Train: 38 502 038 Validation: 20 823 Test: 1997
MuCoW WMT19 Translation Test Suite	https://git.io/Jt858	EN–DE: 3337 EN–RU: 1814
WinoMT	https://github.com/gabrielStanovsky/mt_gender	All: 3888 Without neutral: 3648

Table A2: Details of the datasets used in the paper

E Human Validation

For each language and each task we annotate a subsample of machine translations and compute the proportion of agreement between human evaluation and automated evaluation methods. The annotations are created as follows:

- For MuCoW, we translate the in-domain source sentences with state-of-the-art ensembles (Ng et al., 2019). We first evaluate the translations using the pattern-matching evaluation method proposed by Scherrer et al. (2020), using Stanza (Qi et al., 2020) for lemmatization. We then randomly select a subset of translations for validation: Up to 200 translations that are undecidable given pattern-matching evaluation, and a larger subset of decidable translations proportionate to the overall ratio of decidable translations. For the former we collect human annotations, for the latter we assume that the pattern-matching evaluation is correct.
- With regard to WinoMT, we annotate translations originally collected by Stanovsky et al. (2019)[7]. For English–German, we use 200 translations extracted from Amazon Translate that are a superset of the human validation data used by Stanovsky et al. (2019); for English–Russian we do the same with translations from Google Translate.

In both datasets, annotators have found some edge cases, which we handle as follows when converting the raw labels to binary labels: In MuCoW, we skip some samples that have been marked by our annotators as badly defined, e.g. because the sense definitions overlap too much, or because the gold label is wrong due to a misaligned or mistranslated reference. This only affects the subset of the samples that are undecidable for the original MuCoW algorithm, since we do not manually review the other samples. In WinoMT, we skip samples with a neutral gold label because they are out of the scope of this study (Section 3.6). Furthermore, some translations have been marked as neutral because they preserve the ambiguity of the source (e.g. gender-neutral translations); we treat those cases as correct translations. Finally, evaluators have marked a small number of translations as undecidable, e.g. because the ambiguous part of the input was ignored by the MT system; we treat those cases as disambiguation errors.

Inter-annotator agreement is reported in Table A3. In Table A4 we compare the annotations originally collected by Stanovsky et al. (2019) to the corresponding subset of our own WinoMT annotations. The moderate agreement on WinoMT underlines that especially in Russian, the interpretation of occupation nouns can be subjective.

Based on the human annotations, we compute the proportion of agreement of different evaluation approaches. To ensure a fair comparison with the pattern-matching approach to MuCoW, we do not treat all indecisions as disagreements. Instead we follow the notion of *recall* (B) proposed by Scherrer et al. (2020) and judge undecidable translations as wrong, which may be in agreement or disagreement with the human annotator. For pattern-matching evaluation of WinoMT, we use the reference imple-

[7] https://github.com/gabrielStanovsky/mt_gender

	EN–DE	EN–RU
MuCoW		
– raw labels	0.38	0.57
– binarized labels	0.92	0.88
WinoMT		
– raw labels	0.95	0.13
– binarized labels	0.95	0.20

Table A3: Inter-annotator agreement as measured by Cohen's kappa coefficient. The first line reports the agreement of raw labels, of which there are 4–5 per task (e.g. 'Both'). The second line reports the agreement of the corresponding binary labels after handling the edge cases as described in Appendix E.

	Agreement with original annotations
EN–DE Annotator 1	0.27
EN–DE Annotator 2	0.30
EN–RU Annotator 1	0.57
EN–RU Annotator 2	0.08

Table A4: Agreement between our WinoMT annotators and human annotations by Stanovsky et al. (2019).

mentation and make sure that the word alignment is computed on the full dataset, rather than the validation subset.

Annotator Guidelines for MuCoW

The goal of this annotation is to evaluate machine translation of ambiguous nouns: Is the translation closer to the correct sense cluster than to the wrong one?

Explanation of the data provided to you:

Word The ambiguous source word

Correct Sense The correct sense cluster(s) as represented by a list of German words

Wrong Sense Other sense clusters as represented by a list of German words

Translation The machine-translated sentence that is evaluated

Source Sentence The original English sentence

Explanation of the labels:

Correct Sense The translation is closer to the correct sense cluster.

Wrong Sense The translation is closer to the other sense clusters.

Both / Neutral / Ambiguous The translation preserves the ambiguity found in the source sentence.

Bad sample / Ill-defined senses The sample is not adequate for evaluating word sense disambiguation, e.g. due to overlapping sense clusters or because the gold senses are not consistent with the source sentence.

Translation too bad to tell / Third sense It is impossible to assign a label due to bad translation quality.

Annotator Guidelines for WinoMT

The goal of this annotation is to evaluate machine translation of occupation names: Does the occupation name convey the gender implied by its context?

Explanation of the data provided to you:

Entity The evaluated occupation in English. Please note that only one occupation name per sentence is evaluated, even though the sentence might contain multiple occupations.

Translation The machine-translated sentence that is evaluated.

Source Sentence The original English sentence

Explanation of the labels:

Male The occupation name conveys a male gender.

Female The occupation name conveys a female gender.

Both / Neutral / Ambiguous The translation preserves the ambiguity found in the source sentence.

Translation too bad to tell It is impossible to assign a label due to bad translation quality.

Other remarks:

- Please annotate semantical gender, not grammatical gender.
- Please only take into consideration the occupation noun itself and associated articles. Specifically, try to ignore any pronouns referring to the occupation noun. Pronouns will often disagree with the occupation noun. It is of utmost importance that the pronouns do not influence your annotation. To give an example in English, the label for the following sentence should be 'female', not 'male': *The actress looked at himself in the mirror.*

F Further Examples of Contrastive Conditioning for Word Sense Disambiguation

Example Inputs	Gold Rating / Scores
Ambiguous source: *In light of these findings, we'd like to offer a **settlement**.*	
Machine translation:	
*Angesichts dieser Erkenntnisse möchten wir einen **Vergleich** anbieten.*	correct
– Source with correct disambiguation cue:	
*In light of these findings, we'd like to offer a **cash settlement**.*	0.63
– Source with incorrect disambiguation cue:	
*In light of these findings, we'd like to offer a **military settlement**.*	0.41
Ambiguous source: *There's no van in West Virginia with the 2H7 sequence on its **plate**.*	
Machine translation:	
*In West Virginia gibt es keinen Van mit der 2H7-Sequenz auf dem **Teller**.*	incorrect
– Source with correct disambiguation cue:	
*There's no van in West Virginia with the 2H7 sequence on its **number plate**.*	0.56
– Source with incorrect disambiguation cue:	
*There's no van in West Virginia with the 2H7 sequence on its **dessert plate**.*	0.76
Ambiguous source: *I was dead wrong about that cheetah playing **tag** with the gazelle on Animal Planet.*	
Machine translation:	
Ich lag völlig falsch, als dieser Gepard mit der Gazelle auf Animal Planet spielte.	deleted
– Source with correct disambiguation cue:	
*I was dead wrong about that cheetah playing **freeze tag** with the gazelle on Animal Planet.*	0.69
– Source with incorrect disambiguation cue:	
*I was dead wrong about that cheetah playing **name tag** with the gazelle on Animal Planet.*	0.69
Ambiguous source: *It never used to feel like a **strain**.*	
Machine translation:	
Никогда не было ощущения **напряжения**.	correct
– Source with correct disambiguation cue:	
*It never used to feel like a **stress strain**.*	0.62
– Source with incorrect disambiguation cue:	
*It never used to feel like a **genetic strain**.*	0.33
Ambiguous source: *The **conductor** must be complete for our plan to succeed.*	
Machine translation:	
Дирижер должен быть полным, чтобы наш план увенчался успехом.	incorrect
– Source with correct disambiguation cue:	
*The **light conductor** must be complete for our plan to succeed.*	0.63
– Source with incorrect disambiguation cue:	
*The **choir conductor** must be complete for our plan to succeed.*	0.69

Table A5: Examples of contrastive conditioning based on the MuCoW testset (Raganato et al., 2019) for WSD. For each example, we show the correctly modified source and the incorrectly modified source with the highest evaluator scores. The modifications were generated with a masked language model and selected as described in Appendix D. A higher score means a better score, thus, contrastive conditioning generally agrees with the human gold rating. A special case is the third German example, where the ambiguous word (*tag*) is deleted in the translation. Contrastive conditioning assigns very similar scores to the contrastive sources of that example.

G Further Examples of Contrastive Conditioning for Gendered Occupations

Example Inputs	Gold Rating / Scores
Ambiguous source: *The **farmer** gave the teacher a bag of eggs because he was grateful.*	
Machine translation:	
*Der **Bauer** gab dem Lehrer eine Tüte Eier, weil er dankbar war.*	correct
– Source with correct disambiguation cue:	
*The **[male] farmer** gave the teacher a bag of eggs because he was grateful.*	0.76
– Source with incorrect disambiguation cue:	
*The **[female] farmer** gave the teacher a bag of eggs because he was grateful.*	0.68
Ambiguous source: *The secretary greeted the **supervisor** and handed her the report.*	
Machine translation:	
*Die Sekretärin begrüßte die **Vorgesetzte** und überreichte ihr den Bericht.*	correct
– Source with correct disambiguation cue:	
*The secretary greeted the **[female] supervisor** and handed her the report.*	0.74
– Source with incorrect disambiguation cue:	
*The secretary greeted the **[male] supervisor** and handed her the report.*	0.69
Ambiguous source: *The supervisor gave the **housekeeper** a ride because he lives far away.*	
Machine translation:	
*Der Vorgesetzter hat die **Haushälterin** mitgenommen, weil er weit weg lebt.*	incorrect
– Source with correct disambiguation cue:	
*The supervisor gave the **[male] housekeeper** a ride because he lives far away.*	0.57
– Source with incorrect disambiguation cue:	
*The supervisor gave the **[female] housekeeper** a ride because he lives far away.*	0.59
Ambiguous source: *The receptionist helped the **construction worker** and got a present from her.*	
Machine translation:	
*Der Rezeptionist hat dem **Bauarbeiter** geholfen und hat ein Geschenk von ihr bekommen.*	incorrect
– Source with correct disambiguation cue:	
*The receptionist helped the **[female] construction worker** and got a present from her.*	0.50
– Source with incorrect disambiguation cue:	
*The receptionist helped the **[male] construction worker** and got a present from her.*	0.56
Ambiguous source: *The **guard** asked the librarian to leave because it was late and she will be leaving soon.*	
Machine translation:	
*Die **Wache** bat den Bibliothekar zu gehen, weil es spät war und sie wird bald gehen.*	neutral
– Source with correct disambiguation cue:	
*The **[female] guard** asked the librarian to leave because it was late and she will be leaving soon.*	0.57
– Source with incorrect disambiguation cue:	
*The **[male] guard** asked the librarian to leave because it was late and she will be leaving soon.*	0.56

Table A6: English–German examples of contrastive conditioning based on the WinoMT testset (Stanovsky et al., 2019) for gendered occupation names. Again, contrastive conditioning tends to agree with the gold rating. The final example is a special case, since the translation is gender-neutral and similar scores are assigned to the two contrastive sources.

H Detailed Results for the Case Study

EN–DE Models

Model			newstest19	MuCoW Accuracy			WinoMT Accuracy		
Type	Size	No.	BLEU	total	'freq.'	'rare'	total	'male'	'female'
SOTA Teacher	big	1	40.6	87.2	93.9	79.6	79.1	91.7	66.3
		2	40.8	87.3	94.1	79.7	76.8	89.8	63.8
		3	42.7	86.8	94.1	78.7	70.0	89.5	50.3
Student of SOTA Teacher	big	1	42.1	83.4	93.4	72.1	71.6	90.6	52.5
		2	41.6	82.6	92.9	71.0	70.5	92.6	48.4
		3	42.4	82.0	93.0	69.7	74.4	93.4	55.3
Student of SOTA Teacher	small	1	38.4	76.8	91.1	60.7	68.2	98.1	38.1
		2	38.7	76.8	90.4	61.4	69.3	97.0	41.5
		3	38.8	76.6	90.3	61.2	67.1	97.0	37.2
Student of SOTA Teacher	mini	1	34.5	71.7	86.4	55.2	48.9	97.1	0.6
		2	34.7	71.7	86.1	55.4	49.0	97.4	0.4
		3	34.8	72.3	86.5	56.4	48.8	96.8	0.6
Scratch Teacher	big	1	36.9	79.1	91.8	64.8	63.5	97.3	29.6
		2	37.5	77.7	90.5	63.2	65.5	97.5	33.4
		3	37.4	80.0	92.1	66.5	61.1	96.5	25.6
Scratch Model	small	1	34.1	75.9	89.4	60.7	58.8	96.5	21.0
		2	34.3	75.4	88.4	60.7	55.8	96.6	14.8
		3	33.7	75.7	89.9	59.7	51.7	93.7	9.6
Scratch Model	mini	1	29.8	71.7	84.8	56.9	48.5	96.0	0.8
		2	29.7	71.7	85.2	56.5	47.4	91.4	3.3
		3	29.9	72.1	86.4	56.1	48.7	96.0	1.1
Student of Scratch Teacher	big	1	36.8	78.5	91.7	63.6	63.0	94.9	30.9
		2	37.9	77.1	91.1	61.4	63.1	97.2	28.8
		3	37.6	78.9	92.6	63.5	58.7	93.2	23.9
Student of Scratch Teacher	small	1	35.5	75.2	89.4	59.2	60.9	95.3	26.4
		2	36.1	75.4	89.2	59.9	62.0	97.0	26.8
		3	35.4	75.9	89.1	61.1	54.2	96.9	11.4
Student of Scratch Teacher	mini	1	32.4	71.6	86.3	55.0	56.8	96.0	17.4
		2	33.1	72.0	87.3	54.8	48.3	95.5	0.9
		3	32.3	72.2	86.8	55.7	54.3	96.3	12.2

EN–RU Models

Model			**newstest19**	**MuCoW Accuracy**			**WinoMT Accuracy**		
Type	Size	No.	BLEU	total	'freq.'	'rare'	total	'male'	'female'
SOTA Teacher	big	1	31.1	90.0	91.7	88.2	62.5	84.1	40.9
		2	31.3	90.9	92.1	89.7	62.5	84.4	40.6
		3	33.3	88.7	91.4	85.8	60.8	84.2	37.3
Student of SOTA Teacher	big	1	29.4	85.2	89.2	81.0	53.1	79.8	26.4
		2	29.3	85.2	89.6	80.7	54.4	81.0	27.6
		3	30.0	85.7	90.3	81.0	52.2	81.1	23.2
Student of SOTA Teacher	small	1	25.7	81.7	86.8	76.3	50.4	79.1	21.6
		2	25.7	82.4	86.9	77.7	51.2	81.2	21.2
		3	25.9	82.4	87.1	77.6	49.5	83.4	15.5
Student of SOTA Teacher	mini	1	24.9	78.5	87.4	69.2	51.7	88.3	14.9
		2	24.9	79.1	87.8	70.2	52.0	87.6	16.3
		3	25.0	78.0	87.8	67.8	49.2	86.7	11.5
Scratch Teacher	big	1	26.7	83.3	87.4	79.1	50.3	86.8	13.7
		2	26.2	83.2	88.9	77.2	54.2	86.9	21.3
		3	26.3	83.6	87.5	79.5	53.9	83.1	24.5
Scratch Model	small	1	24.8	81.9	89.0	74.4	49.6	89.6	9.5
		2	24.5	81.9	90.2	73.2	49.8	89.3	10.2
		3	24.6	82.7	89.9	75.1	50.1	89.4	10.6
Scratch Model	mini	1	21.2	79.3	86.8	71.4	48.2	88.2	8.1
		2	21.2	79.7	87.0	72.0	47.8	87.5	8.0
		3	21.1	78.7	86.5	70.5	47.2	84.3	10.0
Student of Scratch Teacher	big	1	28.4	85.7	92.2	79.0	51.4	91.6	11.1
		2	28.4	84.8	90.8	78.5	50.7	91.9	9.4
		3	28.2	84.1	90.4	77.4	49.0	90.2	7.6
Student of Scratch Teacher	small	1	25.4	83.2	91.0	75.1	49.8	89.6	9.9
		2	25.6	82.2	88.9	75.3	49.8	88.7	10.7
		3	25.4	81.7	88.4	74.8	50.2	87.7	12.5
Student of Scratch Teacher	mini	1	22.5	80.1	87.6	72.3	48.5	89.7	7.2
		2	22.5	79.5	87.9	70.8	48.1	87.8	8.2
		3	22.5	78.5	87.4	69.1	48.3	88.7	7.7