Evaluating and Improving the Coreference Capabilities of Machine Translation Models

Asaf Yehudai[1] **Arie Cattan**[2] **Omri Abend**[1] **Gabriel Stanovsky**[1]
[1]School of Computer Science, The Hebrew University of Jerusalem
[2]Computer Science Department, Bar Ilan University
{asaf.yehudai,omri.abend,gabriel.stanovsky}@mail.huji.ac.il
arie.cattan@gmail.com

Abstract

Machine translation (MT) requires a wide range of linguistic capabilities, which current end-to-end models are expected to learn implicitly by observing aligned sentences in bilingual corpora. In this work, we ask: *How well do MT models learn coreference resolution from implicit signal?* To answer this question, we develop an evaluation methodology that derives coreference clusters from MT output and evaluates them without requiring annotations in the target language. We further evaluate several prominent open-source and commercial MT systems, translating from English to six target languages, and compare them to state-of-the-art coreference resolvers on three challenging benchmarks. Our results show that the monolingual resolvers greatly outperform MT models. Motivated by this result, we experiment with different methods for incorporating the output of coreference resolution models in MT, showing improvement over strong baselines.[1]

1 Introduction

Machine translation (MT) may require coreference resolution to translate cases where the source and target language differ in their grammatical properties. For example, consider translating *"The trophy didn't fit in the suitcase because it was too small"* from English to French: *"Le trophée ne rentrait pas dans la valise car elle était trop petite"* (Sakaguchi et al., 2020). In French, suitcase (*"valise"*) is grammatically feminine, and trophy (*"trophée"*) is masculine, while the source-side English does not encode grammatical noun gender. This requires an MT model to infer that *"it"* refers to the suitcase (and not the trophy) to correctly produce the feminine inflection for the phrase *"it was too small"* (*"elle était trop petite"*), whereas an incorrect coreference resolution may produce the masculine inflection (*"il était trop petit"*) corresponding to the trophy.

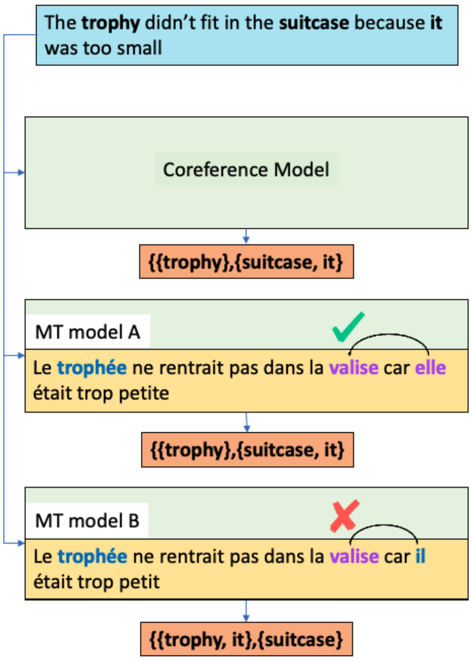
The **trophy** didn't fit in the **suitcase** because **it** was too small
Coreference Model
{{trophy},{suitcase, it}}
MT model A
✓
Le trophée ne rentrait pas dans la valise car elle était trop petite
{{trophy},{suitcase, it}}
MT model B
✗
Le trophée ne rentrait pas dans la valise car il était trop petit
{{trophy, it},{suitcase}}

Figure 1: MT models can be compared to source-side coreference resolvers. An example translation from English (turquoise) to French (yellow). Our method first identifies the grammatical gender of the mentions in the target language marked in purple (female) and blue (male), followed by inferring the source side clusters (orange), through gender agreement.

Such texts evade lexical one-to-one translation, and instead demand source-side coreference resolution as a prerequisite for a correct translation. The prominent end-to-end approach to MT assumes that translation models implicitly learn source-side coreference resolution by observing aligned source-target pairs, without intermediate coreference supervision. While the importance of addressing such semantic phenomenon has been stated in various works (Le Nagard and Koehn, 2010; Stojanovski and Fraser, 2018), it was also observed that the ubiquitous BLEU metric (Papineni et al., 2002) does not adequately quantify it (Hardmeier and Federico, 2010; Freitag et al., 2022).

[1] https://github.com/AsafYehudai/MT-coref

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 980–992
May 2-6, 2023 ©2023 Association for Computational Linguistics

This work addresses the following research question: *How well does MT learn coreference when compared against explicit coreference supervision?* Answering this question can improve our understanding of the way MT models operate and also has practical implications: if implicit supervision lags behind monolingual training, it would motivate integration between end-to-end MT approaches and explicitly-supervised monolingual components.

In Section 3 we devise an evaluation paradigm that reduces MT output to source-side coreference resolution predictions by inferring coreference clusters from source inputs and predicted target translations. E.g., in the previous example, a feminine inflection for the pronoun "it" in French can infer linking "it" with "suitcase", while a masculine French inflection links "it" with "trophy", as shown in Figure 1. This approach allows us to distill the coreference resolution abilities of MT models and compare them against state-of-the-art coreference resolution models, trained explicitly on the task.

We use this approach to evaluate the coreference capabilities of several commercial and open source MT systems, translating from English to six target languages. We conduct our experiments in both synthetic (WinoMT and Wino-X; Stanovsky et al., 2019; Emelin and Sennrich, 2021) and naturalistic settings (BUG; Levy et al., 2021). Our results show that state-of-the-art coreference resolvers vastly outperform MT models on several benchmarks, indicating that explicit supervision may lead to better coreference performance.

Following this finding, in Section 4, we develop methods for improving coreference in MT, both implicitly and explicitly. Our implicit approach consists of fine-tuning MT models on texts that specifically require many coreference decisions, thus exposing the model to more implicit coreference signal. Our explicit approach further enriches source sentences with predicted coreference markers. We show that these approaches improve coreference over the end-to-end MT approach, achieving comparable or better results than much larger MT models, both commercial systems and open-source.

More broadly, our approach can be applied to improve the translation of other semantic phenomena that diverge in realization between source and target languages, such as plurality in second-person pronouns (Stanovsky and Tamari, 2019) or tense marking (Wolfram, 1985).

2 Background: Gender Bias in MT

We start our work by extending the methodology developed in (Stanovsky et al., 2019), which relies on target-side morphology to infer the translated gender of certain professions.

In particular, assuming a dataset of English sentences D, where each instance includes gold coreference annotation between a human entity and its pronoun (e.g., *"The **doctor** asked the nurse to help **her** with the procedure."*), they evaluate gender bias from English to language T with morphological gender in the following manner:

1. Predict word alignment between D and $M(D)$, i.e., the output translations of an MT model M. This finds the translations for pronouns (e.g., *"her"*) and possible entities (e.g *"doctor"*, *"nurse"*) in the target language T.

2. Automatically extract the gender of the possible entities and the pronouns in the target language based on morphological features.

3. Check whether the gender of the co-referred entity (e.g., *"doctor"*) in T corresponds to the gender of the English pronoun (e.g., *"her"*).

The gender bias of M is then defined as the difference in performance between stereotypical and anti-stereotypical gender role assignments.

We use a similar setup to address a different question: rather than evaluating the gender bias of the model, we evaluate its coreference abilities, which may be hindered by bias, but also by the inherent difficulty to infer coreference in the absence of an explicit training signal.

3 MT Models Fare Poorly Against Coreference Resolvers

The approach taken in WinoMT is limited as it restricts the evaluation to sentences with a known gender in English, indicated by a gendered pronoun of a human entity (e.g., *her*). Consider the sentence in Figure 1: *"The trophy didn't fit in the suitcase because it was too small"* with the coreference cluster $\{$suitcase, it$\}$. Step 3 in Stanovsky et al.'s method will fail to assign a coreference label to the translation, because *"it"* does not have a gender in English. In this section, we extend the WinoMT approach in order to estimate more general coreference abilities of MT models.

To achieve this, we note that many languages have gender agreement between pronouns and the noun that they refer to. Therefore, correct target-side gender agreement requires (implicitly) resolving the source-side coreference of the relevant entities. As exemplified in Figure 1, a reader of the French translation would infer from the gender inflection of "it" whether it refers to "suitcase" or the "trophy". I.e., a feminine pronoun ("elle") would agree with the feminine noun suitcase ("valise"), while a masculine pronoun ("il") would agree with the masculine noun trophy ("trophée'). We therefore formulate a new metric quantifying the ability of the MT model to implicitly resolve source-side coreference (henceforth, *Target-side Consistency*), defined as the proportion of instances in which the morphological gender of an entity (e.g., "suitcase") matches that of its referring pronoun (e.g., "it") in the target language T.

This metric examines whether an MT model is consistent in its coreference decisions, regardless of whether it correctly inferred the coreference relations in the input text. Indeed, some texts may keep the English ambiguity in the translation, and hence absolve the MT model from resolving coreference. For example, in the sentence *"The battery didn't fit in the suitcase because it was too small"*, both *battery* and *suitcase* are feminine in French. Our proposed metric will correctly indicate that the MT model was successful in such cases (albeit trivially). The metric thus serves as an upper bound on the MT model's coreference abilities.

We note that while this framing uses the morphological gender inflection of common nouns, it is different in motivation from measures of gender bias. In our example above, gender inflection allows us to determine whether an MT model correctly employs common sense rather than examining whether it tends to prefer stereotypical gender norms. While a model's gender bias may explain some loss in coreference abilities, the model's ability to resolve coreference need not be aligned with the degree of its bias (e.g., a random gender assignment would result in unbiased performance, but very poor coreference ability).

Most importantly, by considering the gender of the entity and the pronoun, we obtain mention clusters which can be compared against those produced by coreference resolution models. In our example figure, both the first MT model and the coreference model produce the correct clustering: {{trophy}, {suitcase, it}}, while the second MT model errs by producing: {{trophy, it}, {suitcase}}.

Another aspect of our evaluation methodology is its generality. Our method does not require a reference translation or make any particular assumptions about the generated output. As there are generally many correct translations, this flexibility allows us to accurately assess the model's coreference abilities. For instance, our methodology does not assume the gender of the entity's translation as can be seen in the first example in Table 7 where the two systems translate the entity "jar" differently. Further, some languages might not always translate an English pronoun into a pronoun but still express its gender in a different word. Consider the second example in Table 7 where the alignment model (step 2 in §2) finds that the English word "it" is aligned with both "l" and "trouvée" in French. Here, the feminine suffix of the past participle "trouv**ée**" indicates that the ellipsis "l" corresponds to a feminine entity.

3.1 Evaluation Setup

Evaluation datasets. The first dataset we use is Wino-X (Emelin and Sennrich, 2021), a filtered subset of WinoGrande (Sakaguchi et al., 2020) built to test commonsense reasoning and coreference resolution of MT models and multilingual encoders. The dataset contains sentences similar to the one in Figure 1. All sentences have two entities and a pronoun, "it", coreferring to one of them. The dataset consists of three parts, each part constructed for a different target language (German, French, and Russian), where each part only contains sentences where the two entities have different morphological gender in the target language (e.g., in French *"trophy"* is masculine whereas *"suitcase"* is feminine). Hence, applying our target-side consistency metric on these filtered sentences avoids trivial instances where both candidate entities have the same gender in the target language, and provides a clearer picture of the coreference capabilities of the model.

Second, we use WinoMT (Stanovsky et al., 2019),[2] a dataset built following the Winograd schema (Levesque et al., 2011), designed to test gender bias and coreference resolution of MT models. The sentences in this dataset contain two human entities and one gendered pronoun, e.g., *"The doctor asked the nurse to help her in the proce-*

[2] WinoMT is a combination of Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018).

Dataset	#sentences
Wino-X (en → de)	3,774
Wino-X (en → fr)	2,988
Wino-X (en → ru)	2,238
WinoMT	3,888
BUG	1,717

Table 1: Statistics of our evaluation datasets. Note that in all datasets we use the corresponding English source-side sentences as our input.

dure". The gendered pronoun reveals the gender of the entity and adds gender attributes to the source cluster. In our example, *"her"* refers to the *"doctor"*, revealing the doctor's gender.

Our third dataset is BUG (Levy et al., 2021), a semi-automatic collection of naturalistic English sentences that are challenging with respect to societal gender-role assignments. Similar to WinoMT, each sentence contains a human entity, identified by their profession and a gendered pronoun. To reduce noise, we use the GOLD portion of this dataset which was validated by human annotators. All datasets statistics are presented in Table 1.

Machine translation models. We apply our evaluation methodology to four Transformer-based machine translation models from EasyNMT:[3] mBART50 (Tang et al., 2020; Liu et al., 2020), M2M_418M, M2M_1.2B (Fan et al., 2021), and the bilingual Opus-MT (Tiedemann and Thottingal, 2020), representing the state-of-the-art for publicly available neural machine translations models. In addition, we measure coreference consistency on the output of two commercial systems: Google Translate[4] and Microsoft Translator.[5]

Target languages. For WinoMT and BUG, we translate from English to six different languages: Arabic, German, Spanish, Hebrew, Russian and French. These languages form a diverse set with respect to how they encode grammatical gender (e.g., number of grammatical genders), as well as to their orthography, word order and other linguistic traits, while still allowing for highly accurate automatic morphological analysis. These languages belong to four families: (1) Romance languages: Spanish and French, which have gendered noun-determiner agreement with two grammatical genders; Spanish is also a *pro-drop* language, i.e., pronouns can be omitted in certain cases, which in our setting may keep the coreference ambiguity of the source-side English sentence (Webster and Pitler, 2020). (2) Slavic languages (with Cyrillic alphabets): Russian with 3 grammatical genders. (3) Semitic languages: Hebrew and Arabic, each with a unique alphabet; both are partial pro-drop languages and have two grammatical genders. (4) Germanic languages: German with 3 grammatical genders.

	Wino-X	WinoMT	BUG
SpanBERT	51.2	76.6	72.0
s2e	**60.8**	81.7	72.2
LINGMESS	58.7	**83.7**	**74.6**

Table 2: Accuracy of SpanBERT, *s2e* and LINGMESS model on our evaluation datasets. For simplicity, we report the accuracy on Wino-X sentences from the three languages as a single corpus, because there is a small difference between the languages (up to 0.3).

	en → de	en → fr	en → ru
mBART50	37.4	**56.6**	44.6
M2M_418M	31.4	48.5	**44.9**
Google	**41.3**	36.3	40.7
Microsoft	40.5	36.7	43.5
Opus-MT	37.4	35.3	43.6

Table 3: Target-side consistency results of commercial and open-source MT systems on Wino-X when translating into German, French, and Russian.

3.2 Target-side Consistency Results

We first evaluate the accuracy of existing coreference resolvers on our three evaluation datasets, where accuracy is defined as the percentage of instances in which the model identifies that the pronoun is coreferring with the correct entity. We select state-of-the-art models trained on CoNLL-2012 (Pradhan et al., 2012): SpanBERT (Joshi et al., 2020),[6] the s2e model (Kirstain et al., 2021) and LINGMESS (Otmazgin et al., 2022). Results in Table 2 show that coreference models perform quite well on WinoMT and BUG but poorly on Wino-X (60.8 for *s2e*), indicating weak commonsense capabilities.

Table 3 shows the target-side coreference consistency scores for all MT models on Wino-X, which, as mentioned above (§3.1), includes only sentences

[3] https://github.com/UKPLab/EasyNMT

[4] https://cloud.google.com/translate

[5] https://www.bing.com/translator

[6] Using AllenNLP's implementation (Gardner et al., 2018).

	WinoMT						BUG					
	de	es	fr	ru	he	ar	de	es	fr	ru	he	ar
mBART50	77.7	**75.2**	**73.3**	57.7	**69.3**	**69.5**	74.6	69.4	76.0	69.1	80.2	83.0
M2M_418M	69.7	55.1	65.4	54.7	56.5	64.3	81.3	84.9	86.0	**71.6**	83.6	85.1
M2M_1.2B	69.7	55.1	65.4	54.7	56.5	64.3	**83.5**	85.1	**90.3**	70.4	**89.6**	**93.4**
Google	69.9	59.4	65.5	**57.9**	60.1	69.2	56.0	84.4	89.2	71.0	85.2	91.9
Microsoft	**78.0**	66.5	69.3	57.3	63.9	60.7	77.2	**86.7**	86.0	67.2	84.0	89.8
Opus-MT	68.2	56.0	63.2	49.8	59.0	59.8	79.8	85.1	86.2	67.3	88.8	88.0

Table 4: Target-side consistency results of commercial and open-source MT systems on WinoMT and BUG when translating into different languages. These numbers are an upper bound for the source-side coreference accuracy.

where the entity and pronoun should be translated using different genders in the target language. We observe that all MT models perform poorly on Wino-X with the highest average score of 46.2 for mBART50, which vastly underperforms English coreference resolvers by 14.6 points. Interestingly, many instances are inconsistent because models tend to generate *neutral* pronoun whereas a gendered pronoun is expected, for example, *cela*, *c'était* in French (31% of translations) and это,они in Russian (17% of translations), meaning "this" or "they" in English. Likewise, 68% of German translations include neutral pronouns (e.g., "es"), while only 22% of the entities are neutral. The reported percentages were calculated on Opus-MT. Similar trends were observed in all models.

Table 4 shows the target-side consistency results for WinoMT and BUG. Following common practices on those datasets (Stanovsky et al., 2019; Levy et al., 2021), we omit sentences where the candidate pronoun does not provide information about the entity's gender. For example, in French, possessive pronouns agree with the gender of the possessed object, rather than the possessor as in English. Another example is in Spanish, which is a pro-drop language, where a valid translation can drop the pronoun and use a generic verb, leaving the only gender signal in the translation to be marked on the profession noun. See App. §C for more examples.

Similarly to Wino-X, target-side consistency results on WinoMT are consistently lower than coreference resolvers. Further, we observe that consistency is affected by two factors: the MT model and the target language. Regarding models, Opus-MT achieves lowest performance, with average consistency of 59.3, while mBART50 achieves high results with average consistency of 70.5, sometimes surpassing the second-best MT model by about 9 points. This might be due to the extensive pretraining of mBART50, as previously demonstrated for monolingual LMs (Huang et al., 2019; Sakaguchi et al., 2020; Bhagavatula et al., 2020). With respect to target languages, Russian consistency results are systematically lower than the results in other languages, to the extent that the best model in Russian provides lower results than the worst model in most other languages. In contrast, all models in German achieve a consistency score of about 70 or more, which can be due to its similarity with English and the research focus on improving English-German translations.

Consistency results on BUG are higher than on WinoMT for most models, while sometimes surpassing English coreference resolvers, notably in Hebrew and Arabic (e.g., 91.8 for Google vs. 74.6 for LingMess). To understand this gap, we analyze the translation of 50 BUG sentences to Hebrew and French and find that most instances (45 in Hebrew and 33 in French) do not include a distracting entity which should be translated to a different gender in the target language. As mentioned above (§3), our metric trivially indicates those examples as consistent.

Overall, target-side consistency results across all datasets demonstrate that both open-source and commercial MT systems exhibit rather poor coreference capabilities compared to English coreference models.

3.3 Human Validation

The use of automatic tools in the proposed methodology inevitably implies the introduction of noise into the process. To assess the quality of our measurements, we randomly sampled 50 translations of the Opus-MT model from all evaluation datasets and in all target languages (for a total of 750 annotations), annotating each sample in-house by a native speaker of the target language. The human annotators were asked to identify if the candidate pronoun is indeed the target pronoun and to verify that the gender prediction is correct. This way, we can account for both types of possible errors, i.e., alignment and gender extraction.

We compare the human annotations to the output of our automatic method and find that the average agreement over all languages and datasets is above 90% (see full results in App. §A). These results are comparable to the ones reported by Stanovsky et al. (2019), who conducted human validation and reported that their alignment and gender prediction of the entity in question were reliable for 85% of translations across all languages.

Some errors can be caused by idiosyncrasies that affect the morphological analysis, as Stanovsky et al. (2019) noted. For example, gender for certain words in Hebrew cannot be determined without diacritics, and some pronouns in German are used in both masculine and neutral forms (e.g., sein), or feminine and third-person plural forms (e.g., ihr). In addition, we notice that sentences from BUG, specifically in partial pro-drop languages, were found to be more challenging for the alignment model, and account for most mistakes in Hebrew and Arabic.

4 Improving MT Coreference Consistency

In the previous section we showed that the coreference performance of MT systems, obtained through an implicit signal, seems inferior to that of coreference resolution learned from an explicit signal. This result raises the question of whether we can leverage dedicated conference resolvers to improve the consistency of MT coreference.

To address this question, we propose two data augmentation techniques that leverage a source-side English coreference model, and show that fine-tuning on them indeed improves coreference resolution in MT.

Augmented fine-tuning with instances which require coreference resolution. First, we run a coreference resolution model on the source-side sentences. We then consider two approaches for constructing the augmented fine-tuning data: (1) *Coref data* with all sentences that have non-singleton clusters and (2) *Gender data*, a subset of *Coref data* where there is at least one non-singleton cluster with a gendered pronoun (he, she, her, him, hers, his). The motivation for this augmented fine-tuning strategy is that further fine-tuning on such instances would expose the MT model to examples that may bear a coreference signal.

	de	es	fr	ru	he	ar
Coref data	500K	744K	761K	149K	1.1M	1.4M
Gender data	38K	50K	51K	19K	265K	268K

Table 5: Number of fine-tuning instances in *Coref Data* (requiring some sort of coreference resolution) and *Gender data* (requiring coreference resolution with some gendered pronoun) for each target language.

Adding explicit source-side coreference markers. Second, we use the non-singleton clusters from the coreference model to add inline coreference markers in the source sentences. For our example sentence, this process produces the following source-side sequence: *"The trophy didn't fit in the* ***<ENT1>*** *suitcase* ***</ENT1>*** *because* ***<ENT1>*** *it* ***</ENT1>*** *was too small"*, indicating that "suitcase" and "it" are coreferring.

4.1 Experimental Setup

MT models. In our fine-tuning experiments, we opt for the Opus-MT model, since its size (68M parameters) and efficiency (Junczys-Dowmunt et al., 2018) enables us to run extensive experiments across many languages.

Training datasets. For fine-tuning data of Spanish, French, and German, we use Europarl (Koehn, 2005), and for Russian, Hebrew, and Arabic, we use CCMatrix (Schwenk et al., 2021), randomly sub-sampled to 5M sentences for computational reasons. In each dataset, we find instances that require coreference resolution and add appropriate markup using the *s2e* coreference resolver. We use *s2e* as it performed well in the previous experiments. Table 5 shows the size of *Coref data* and *Gender data* for all training datasets. Note that invariably *Gender data* is an order of magnitude smaller than *Coref data*.

Fine-tuning and inference. For each language, we fine-tune the Opus-MT model using four different finetuning datasets: (1) *Coref data* (2) *Coref data* with explicit coreference markers, (3) *Gender data* and (4) *Gender data* with explicit coreference markers. The inference on our three evaluation datasets (Wino-X, WinoMT, BUG) conforms with the fine-tuning procedure of each model. Namely, we run the models (1) and (3) on raw English sentences. For the models (2) and (4), we first add explicit coreference markers according to the output of the *s2e* model (a) or the gold annotation (b), then translate those augmented sentences to the different languages.

4.2 Results

Table 6 shows the target-side coreference consistency scores of all our fine-tuned models on Wino-X and WinoMT (see App. §B for the performance on BUG, which follow similar trends). For both datasets, our fine-tuned models surpass the Opus-MT baseline model, while preserving the overall translation quality, as indicated by automatic measures such as BERTScore (Zhang et al., 2020) (+0.08%) and COMET-20 (Rei et al., 2020) (+0.0025).

Effect of augmented fine-tuning data. The models fine-tuned on *Coref data* (1) and *Gender data* (3) outperform the Opus-MT baseline for all languages, both in Wino-X and WinoMT. This demonstrates that MT models learn implicitly linguistic phenomena from instances involving those phenomena. Furthermore, we point out that consistency scores on Wino-X are generally higher when fine-tuning on *Coref data* (1, 2a, 2b) while WinoMT results are better when fine-tuning on *Gender data* (3, 4a, 4b). This performance gap likely stems from the similarity between WinoMT and *Gender data* (as both include gendered pronouns), while Wino-X's like sentences with the pronoun "it" appear only in *Coref data*. This further confirms the important role of fine-tuning data, which is in line with the observation of Saunders and Byrne (2020), that smaller, more goal-oriented data is better for fine-tuning, compared to much larger but less focused data.

Effect of explicit coreference markers. In the majority of our experiments (13/18), the explicit fine-tuning models (2a and 4a) outperform the implicit data augmentation approach when using the same augmented data (1 and 3) (see examples in Table 7). These results suggest that an explicit monolingual signal can improve results more than an implicit signal. Results also show that the improvement is more pronounced when incorporating gold coreference markers (2b and 4b) instead of predicted markers (2a and 2b). Hence, applying more accurate coreference resolution models than the *s2e* model will result in higher target-side consistency results.

4.3 Analysis

We turn to observing the empirical effect of the suggested fine-tuning strategies, using additional metrics. For each sentence in Wino-X, we have the gold target pronoun that should appear in its translation. We use it to compute pronoun translation accuracy by comparing the candidate pronoun with the gold target pronoun. Table 8 presents the results. We can see that our method provides a large improvement over the baseline. Comparing these results against those of prominent open-source and commercial MT (see App. §D) shows that our approach outperforms other MT models in German and Russian, and is only second in French.

In WinoMT, Stanovsky et al. (2019) computed *gender accuracy* as the percentage of instances in which the translation preserved the gender of the entity from the original English sentence (§2). Table 9 shows that our approach improved gender accuracy results across all languages except Arabic.

Other metrics that Stanovsky et al. (2019) used, are ΔS and ΔG. ΔS measures the difference in gender accuracy between stereotypical and non-stereotypical gender role assignments (as defined by Zhao et al., 2017), and ΔG measures the difference in performance (F1 score) between source sentences with male and female entities. Our method decreases biases in both ΔS and ΔG by 5-6 points on average, indicating that the explicit signal helps the model in associating the pronoun with the coreferring entity, even in the presence of social and gender biases.

5 Related Work

The study of coreference has a long tradition in machine translation. A long line of work uses pronoun translation as a way of measuring coreference, since BLEU-based evaluation was shown to be insufficient for measuring improvement in coreference (Hardmeier and Federico, 2010).

An alternative evaluation methodology is using automatic reference-based methods that produce a score based on word alignment between the source, reference translation, and translation output, and identification of pronouns in them, such as AutoPRF (Hardmeier and Federico, 2010) and APT (Miculicich Werlen and Popescu-Belis, 2017). Nevertheless, a later human meta-evaluation showed

	Wino-X			WinoMT					
	en→de	en→fr	en→ru	en→de	en→es	en→fr	en→ru	en→he	en→ar
Opus-MT	37.4	35.3	43.6	67.6	56.0	63.2	49.8	59.0	59.8
Coref data (1)	42.7	41.0	44.0	74.4	58.0	67.1	**62.1**	68.1	67.1
+coref markers (2a)	44.1	42.6	45.5	76.0	58.2	67.2	57.7	68.7	67.9
+gold markers (2b)	**44.7**	**45.8**	**50.5**	77.6	58.2	67.3	58.6	69.8	**68.9**
Gender data (3)	41.2	37.0	45.0	75.0	60.6	68.3	61.6	68.1	66.2
+coref markers (4a)	43.6	36.6	43.8	78.7	60.1	**68.6**	58.4	69.0	60.9
+gold markers (4b)	42.9	36.7	46.3	**80.6**	**60.8**	68.4	59.6	**69.9**	62.2

Table 6: Target-side consistency results of the Opus-MT baseline and our fine-tuning experiments on Wino-X and WinoMT when translating into different languages. For both datasets, our fine-tuned models surpass the baseline.

Source	The chef tried to store the **fat** in the **jar** but **it** was too large.	
Baseline (FR)	Le chef a essayé de stocker la graisse dans le bocal, mais il était trop grand.	✗
Ours (FR)	Le chef a essayé de stocker la graisse dans le pot, mais elle était trop grande.	✓
Source	The chickens escaped from the **yard** and fled to the **field**, as they found **it** so confining.	
Baseline (FR)	Les poulets se sont échappés de la cour et ont fui vers le champ, comme ils l'ont trouvé si restreint.	✗
Ours (FR)	Les poulets se sont échappés de la cour et ont fui vers le champ, car ils l'ont trouvée si encombrée.	✓
Source	The headphones blocked the **noise** but not the **vibration**, as **it** was relatively strong	
Baseline (RU)	Наушники блокировали шум, но не вибрацию, поскольку он был относительно сильным.	✗
Ours (RU)	Наушники блокировали шум, но не вибрацию, так как она была относительно сильной.	✓

Table 7: Translation examples of Wino-X sentences to French and Russian by the baseline (Opus-MT) and our model (with coref markers). Words in blue and red indicate male, female entities, respectively. Bold indicates coreference mentions in the source sentence.

	en→de	en→fr	en→ru
Opus-MT	39.8	31.7	37.0
Coref data	42.3	38.9	35.0
+coref markers	43.6	39.3	37.2
+gold markers	**45.7**	**42.7**	**42.6**

Table 8: Pronoun accuracy results of our fine tuning approaches on Wino-X.

	de	es	fr	ru	he	ar
Opus-MT	66	60.4	56.5	50.2	56.6	**59.8**
Gender data	73.3	66.7	**60.3**	52.4	60.9	59.7
+coref markers	**76.8**	**67.9**	60.1	**53.2**	**63.5**	55.6

Table 9: Gender Accuracy results of our fine tuning approaches on WinoMT.

substantial disagreement between these metrics and human annotators, especially because of the existence of valid alternative translations and pronouns than the ones used in the reference (Guillou and Hardmeier, 2018). Based on these conclusions, Sennrich (2017) developed a scoring-based evaluation approach that compares model scores of a predefined set of correct and incorrect translations and evaluates how often the model selects the correct option.

Our method extends (Stanovsky et al., 2019), which used a reference-free approach by aligning the source and candidate translation, but focused on entity translation accuracy to evaluate gender bias in MT models. The availability of references was assumed by most previous work (Guillou and Hardmeier, 2016; Bawden et al., 2018; Müller et al., 2018; Stojanovski et al., 2020; Emelin and Sennrich, 2021), where most of them are limited to a single language pair. The flexibility afforded by a reference-free approach allows us to evaluate any target language for which an alignment model and morphological analyzer are available. Moreover, our approach is not restricted by a predefined set of translations and can also correctly detect valid translations that are different from the reference.

Several previous methods aimed to improve the coreference abilities of MT models and reduce undesirable biases, by modifying the training data in ways that share some similarities with our method. Vanmassenhove et al. (2018) incorporate a "speaker gender" tag into training data, allowing gender to be conveyed at the sentence level. Similarly, Moryossef et al. (2019) added a prefix to disambiguate the coreference in the sentence. Stojanovski and Fraser (2018) used oracle-based approach to inject new tokens indicating the pronoun translation and its gender into the source sentence. Our method is novel in the way it enriches the data with coreference signal using only the source-side signal, and thus requires only an English coreference resolution model without the need for coreference annotation in the target language.

6 Conclusion

Our work is the first to present an automatic methodology for assessing the coreference capabilities of MT models, that can be applied in any target language and does not require any target side annotations. Furthermore, to the best of our knowledge, we are the first to conduct a large-scale multilingual coreference evaluation study on prominent open-source and commercial MT models, and compare them against state-of-the-art coreference resolvers on three challenging benchmarks. Finally, based on the superior results of coreference resolvers, we propose a novel approach to improve the coreference capabilities of MT models, that outperforms or achieves comparable results to strong and larger MT models. Despite this substantial gain, there is still a performance gap between our model and state-of-the-art coreference resolvers. We hope that our work, and specifically our automatic evaluation methodology, will encourage future research to improve the coreference capabilities of MT models.

Future work can expand our approach to account for number and person agreement phenomena, investigate how to extend our approach to more coreference clusters and more mentions per cluster in intra-sentential as well as inter-sentential settings. Moreover, we intend to investigate how different morphological attributes affect MT models' coreference abilities.

Limitations

Even though our study presents the first large-scale multilingual coreference evaluation study in MT, it still has some limitations that could be addressed in future work. First, our methodology provides an upper bound to the coreference capabilities based on detecting gender valuations. While this could allow for a controlled evaluation experiment, this upper bound can become non-indicative in cases where gender assignment is not a discriminative factor. This can be addressed by accounting for more semantic and syntactic constraints that the translation needs to follow (e.g., singular/plural agreement).

Second, our setting addresses one entity and a single co-referring pronoun in the naturalistic sentences experiment. Our methodology could in principle be augmented to deal with more coreference clusters and mentions per cluster. Another possible extension is to include event coreference in addition to entity coreference. For example, in this work, we focus only on the anaphoric function of the pronoun "it" but further research can also examine the event function of "it" (Loáiciga et al., 2017).

Third, MT models should generally produce translations with accurate gender inflection for all words. However, in this work, we focus on the *coreference* capabilities of MT models by evaluating gender agreement between coreferring *entity* mentions. Future research can extend our evaluation methodology to assess the gender inflection of verb and adjective translation (e.g., the gender of "big" and "small" in Figure 1), using additional tools and resources such as a semantic role labeling model and a dependency parser.

Finally, although in Section 4 we show big gains from the fine-tuning approach, it is clear that there is much room for improving the coreference capabilities of MT models, especially with regard to the performance of state-of-the-art coreference resolvers. We hope this work will help others develop MT models with better coreference capabilities.

Acknowledgements

We thank the reviewers for their insightful comments and suggestions. This work was partially supported by the Israeli Ministry of Science and Technology (grant no. 2088). Arie Cattan is partially supported by the PBC fellowship for outstanding PhD candidates in data science.

References

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.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya

Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *ArXiv*, abs/1908.05739.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi. Association for Computational Linguistics*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ArXiv*, abs/1803.07640.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 283–289, Paris, France.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *ArXiv*, abs/1909.00277.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *ACL*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MTSUMMIT*.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *KR*.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331, Copenhagen, Denmark. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias* *in Natural Language Processing*, pages 49–54, Florence, Italy. 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.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. Lingmess: Linguistically informed multi expert scorers for coreference resolution. *ArXiv*, abs/2205.12644.

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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641.

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.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *ACL*.

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.

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.

Gabriel Stanovsky and Ronen Tamari. 2019. Y'all should read this! identifying plurality in second-person personal pronouns in english texts. In *EMNLP*.

Dario Stojanovski and Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Kellie Webster and Emily Pitler. 2020. Scalable cross lingual pivots to model pronoun gender for translation. *ArXiv*, abs/2006.08881.

Walt Wolfram. 1985. Variability in tense marking: A case for the obvious. *Language Learning*, 35(2):229–253.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020,* *Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

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.

	he	ar	es	ru	fr	de
alignment error	14	19	4	4	4	0
gender prediction error	7	7	0	7	2	2
correct annotation	79	74	96	139	144	148
Total # of annotation	100	100	100	150	150	150

Table 10: Human validation results on our three evaluation datasets and six target languages.

A Human Validation Results

Table 10 shows the complete human annotations results. The results indicate that alignment and gender prediction are accurate in most languages. In Arabic and Hebrew, the alignment error occurs more. A possible explanation for that can be the fact that both those languages are partial pro-drop languages. To verify that those results will not affect our measurement, we verified that the error has similar consistency distributions as the rest of our results.

B BUG Consistency Results

Table 11 presents the target-side consistency results of the Opus-MT baseline model and all our fine tuning approach on BUG. Similarly to Wino-X and WinoMT, our fine-tuned models outperform the baseline.

C Omitted Cases

Table 12 shows translation examples from English to French and Spanish that demonstrate unique features in each language. The first example shows a French translation that contains a possessive pronoun, which does not indicate the gender of the possessor. The second example shows a Spanish translation where the pronoun is omitted. In both cases, we can obtain a correct translation without information concerning the aligned pronoun gender, we therefore exclude them from the evaluation.

D Pronoun Translation Accuracy

Table 13 shows pronoun accuracy results of our baselines on Wino-X. We can notice those results are similar to our consistency results although our methodology does not use any annotated target data. Moreover, those results clearly show that MT models struggle with translating sentences that demand solving the source side coreference resolution.

	en→de	en→es	en→fr	en→ru	en→he	en→ar
Opus-MT	79.8	85.1	86.2	67.3	88.8	88.0
gender data	84.1	**87.4**	88.7	69.9	**90.2**	91.3
+coref markers	**84.2**	**87.4**	**88.8**	**70.1**	89.8	**91.8**

Table 11: Target-side consistency results of our implicit and explicit fine tuning approaches on BUG.

Source	[Target lang.] Predicted translation	Phenomenon
The developer argued with the designer because his idea cannot be implemented.	[FR.] Le développeur a argumenté avec le concepteur parce que son idée ne peut pas être mis en œuvre.	"son" is male because the possessed noun ("idée") is male.
The doctor asked the nurse to help her in the procedure	[ES.] El doctor le pidio a la enfermera que le ayudara con el procedimiento	In Spanish, the pronoun "her" is dropped in the translation.

Table 12: Examples of omitted sentences from our evaluation datasets and their translations. Words in blue and red indicate male and female entities, respectively.

	German	French	Russian
mBART50	38.1	**47.2**	34.7
M2M00_418M	33.8	39.5	32.4
M2M100_1.2B	38.5	39.0	33.7
Google	**44.2**	34.8	35.8
Microsoft	43.1	35.7	**37.8**
EasyNMT	39.9	31.8	37.0

Table 13: Pronoun accuracy results of commercial and open-source MT models on Wino-X.

E Computing Infrastructure

We fine tuned our models using 4 NVIDIA GTX Titan Black GPUs. The run time of the models varies between one hour to 24 hours depending on dataset size.