ContraCAT: Contrastive Coreference Analytical Templates for Machine Translation

Dario Stojanovski[†*], **Benno Krojer**[†*], **Denis Peskov**[‡*], **Alexander Fraser**[†]
[†]Center for Information and Language Processing, LMU Munich
[‡]Computer Science, University of Maryland
{stojanovski,fraser}@cis.lmu.de, benno.krojer@gmail.com,
dpeskov@cs.umd.edu

Abstract

Recent high scores on pronoun translation using context-aware neural machine translation have suggested that current approaches work well. ContraPro is a notable example of a contrastive challenge set for English→German pronoun translation. The high scores achieved by transformer models may suggest that they are able to effectively model the complicated set of inferences required to carry out pronoun translation. This entails the ability to determine which entities could be referred to, identify which entity a source-language pronoun refers to (if any), and access the target-language grammatical gender for that entity. We first show through a series of targeted adversarial attacks that in fact current approaches are not able to model all of this information well. Inserting small amounts of distracting information is enough to strongly reduce scores, which should not be the case. We then create a new template test set ContraCAT, designed to individually assess the ability to handle the specific steps necessary for successful pronoun translation. Our analyses show that current approaches to context-aware NMT rely on a set of surface heuristics, which break down when translations require real reasoning. We also propose an approach for augmenting the training data, with some improvements.

1 Introduction

Machine translation is a complex task which requires diverse linguistic knowledge. The seemingly straightforward translation of the English pronoun *it* into German requires knowledge at the syntactic, discourse and world knowledge levels for proper pronoun coreference resolution (CR). The German third person pronoun can have three genders, determined by its antecedent: masculine (*er*), feminine (*sie*) and neuter (*es*). Previous work (Hardmeier and Federico, 2010; Miculicich Werlen and Popescu-Belis, 2017; Müller et al., 2018) proposed evaluation methods for pronoun translation. This has been of special interest in context-aware NMT models that are capable of using discourse-level information. Despite promising results (Bawden et al., 2018; Müller et al., 2018; Lopes et al., 2020), the question remains: Are transformers (Vaswani et al., 2017) truly *learning* this task, or are they exploiting simple heuristics to make a coreference prediction?

To empirically answer this question, we extend ContraPro (Müller et al., 2018)—a contrastive challenge set for automatic English→German pronoun translation evaluation—by making small adversarial changes in the contextual sentences. Our adversarial attacks on ContraPro show that context-aware transformer NMT models can easily be misled by simple and unimportant changes to the input. However, interpreting the results obtained from adversarial attacks can be difficult. The results indicate that NMT uses brittle heuristics to solve CR, but it is not clear what those heuristics are. In general, it is challenging to design attacks based on modifying ContraPro that can test specific phenomena that may be of interest.

[*]Equal contribution.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4732–4749
Barcelona, Spain (Online), December 8-13, 2020

Start:	The cat and the actor were hungry.
Original sentence	It (?) was hungrier.
Step 1:	The **cat** and the **actor** were hungry.
Markable Detection	It (?) was hungrier.
Step 2:	The cat and the actor were hungry.
Coreference Resolution	**It** (😺) was hungrier.
Step 3:	Der Schauspieler und die Katze waren hungrig.
Language Translation	Er / **Sie** (😺) / Es war hungriger.

Table 1: A hypothetical CR pipeline that sequentially resolves and translates a pronoun.

For this reason, we propose an *independent* set of templates for coreferential pronoun translation evaluation to systematically investigate which heuristics are being used. Inspired by previous work on CR (Raghunathan et al., 2010; Lee et al., 2011), we create a number of templates tailored to evaluating the specific steps of an idealized CR pipeline. We call this collection ContraCAT (😺), **Contra**stive **C**oreference **A**nalytical **T**emplates. The templates are constructed in a completely controlled manner, enabling us to easily create large number of coherent test examples and provide strong conclusions about the CR capabilities of NMT. The procedure we used in creating the templates can be adapted to many language pairs with little effort. Our 😺 results suggest that transformer models do not learn each step of a hypothetical CR pipeline.

We also present a simple data augmentation approach specifically tailored to pronoun translation. The experimental results show that this approach improves scores and robustness on some of our metrics, but it does not fundamentally change the way CR is being handled by NMT.

We publicly release ContraCAT and the adversarial modifications to ContraPro[1].

2 Coreference Resolution in Machine Translation

Addressing discourse phenomena is important for high-quality MT. Apart from document-level coherence and cohesion, anaphoric pronoun translation has proven to be an important testing ground for the ability of context-aware NMT to model discourse. Anaphoric pronoun translation is the focus of several works in context-aware NMT (Bawden et al., 2018; Voita et al., 2018; Stojanovski and Fraser, 2019; Miculicich et al., 2018; Voita et al., 2019; Maruf et al., 2019).

However, the choice of an evaluation metric for CR is nontrivial. BLEU-based evaluation is insufficient for measuring improvement in CR (Hardmeier, 2012) without carefully selecting or modifying test sentences for pronoun translation (Voita et al., 2018; Stojanovski and Fraser, 2018). Alternatives to BLEU include F_1, partial credit, and oracle-guided approaches (Hardmeier and Federico, 2010; Guillou and Hardmeier, 2016; Miculicich Werlen and Popescu-Belis, 2017). However, Guillou and Hardmeier (2018) show that these metrics can miss important cases and propose semi-automatic evaluation. In contrast, our evaluation is *completely* automatic.

We focus on scoring-based evaluation (Sennrich, 2017), which works by creating contrasting pairs and comparing model scores. Accuracy is calculated as how often the model chooses the correct translation from a pool of alternative incorrect translations. Bawden et al. (2018) manually create such a contrastive challenge set for English→French pronoun translation. ContraPro (Müller et al., 2018) follows this work, but creates the challenge set in an automatic way.

We show that making small variations in ContraPro substantially changes the scores. Our work is related to adversarial datasets for testing robustness used in Natural Language Processing tasks such as studying gender bias (Zhao et al., 2018; Rudinger et al., 2018; Stanovsky et al., 2019), natural language inference (Glockner et al., 2018) and classification (Wang et al., 2019).

Jwalapuram et al. (2019) propose a model for pronoun translation evaluation trained on pairs of sentences consisting of the reference and a system output with differing pronouns. However, as Guillou and Hardmeier (2018) point out, this fails to take into account that often there is not

[1] http://cistern.cis.lmu.de/contracat

a 1:1 correspondence between pronouns in different languages. As a result, a system translation may be correct despite not containing the exact pronoun in the reference, and incorrect even if containing the pronoun in the reference, because of differences in the translation of the referent. Moreover, introducing a separate model which needs to be trained before evaluation adds an extra layer of complexity in the evaluation setup and makes interpretability more difficult. In contrast, templates can easily be used to pinpoint specific issues of an NMT model. Our templates follow previous work (Ribeiro et al., 2018; McCoy et al., 2019; Ribeiro et al., 2020) where similar tests are proposed for diagnosing NLP models.

3 Do Androids Dream of Coreference Translation Pipelines?

Imagine a hypothetical coreference pipeline that generates a pronoun in a target language, as illustrated in Table 1. **First**, markables (entities that can be referred to by pronouns) are tagged in the source sentence (we restrict ourselves to concrete entities as we wish to detect gender). Then, the subset of animate entities are detected, and human entities are separated from other animate ones (since *it* cannot refer to a human entity). **Second**, coreferences are resolved in the source language. This entails handling phenomena such as world knowledge, pleonastic *it*, and event references. **Third**, the pronoun is translated into the target language. This requires selecting the correct gender given the referent (if there is one), and selecting the correct grammatical case for the target context (e.g., accusative, if the pronoun is the grammatical object in the target language sentence).

This idealized pipeline would produce the correct pronoun in the target language. The coreference steps resemble the rule-based approach implemented in Stanford CoreNLP's Coref-Annotator (Raghunathan et al., 2010; Lee et al., 2011). However, NMT models are currently unable to decouple the individual steps of this pipeline. We propose to isolate each of these steps through targeted examples.

4 Model

We use a transformer model for all experiments and train a sentence-level model as a baseline. The context-aware model in our experimental setup is a concatenation model (Tiedemann and Scherrer, 2017) (CONCAT) which is trained on a concatenation of consecutive sentences. CONCAT is a standard transformer model and it differs from the sentence-level model only in the way that the training data is supplied to it. The training examples for this model are modified by prepending the previous source and target sentence to the main source and target sentence, respectively. The previous sentence is separated from the main sentence with a special token <SEP>, on both the source and target side. This also applies to how we prepare the ContraPro and ContraCAT data. We train the concatenation model on OpenSubtitles2018 data prepared in this way. We remove documents overlapping with ContraPro. Preprocessing details and model hyper-parameters are presented in the Appendix.

5 Adversarial Attacks

5.1 About ContraPro

ContraPro is a contrastive challenge set for English→German pronoun translation evaluation. The set consists of English sentences containing an anaphoric pronoun "it" and the corresponding German translations. It contains three contrastive translations, differing based on the gender of the translation of *it*: *er*, *sie*, or *es*. The challenge set artificially balances the amount of sentences where *it* is translated to each of these three German pronouns. The appropriate antecedent may be in the main sentence or in a previous sentence. For evaluation, a model needs to produce scores for all three possible translations, which are compared against ContraPro's gold labels.

We create automatic adversarial attacks on ContraPro that modify theoretically inconsequential parts of the context sentence before the occurrence of *it*. Contrary to expectations, we find that accuracy degrades in all adversarial attacks. Results are presented in Figure 1.

5.2 Adversarial Attack Generation

Our three modifications are:

1. **Phrase Addition**: Appending and prepending phrases containing implausible antecedents: The Church is merciful *but that's not the point*. It always welcomes the misguided lamb.

2. **Possessive Extension**: Extending the original antecedent with a possessive noun phrase: I hear ~~her~~ *the doctor's* voice! It resounds to me from heights and chasms a thousand times!

3. **Synonym Replacement**: Replacing the original German antecedent with a synonym of a different gender (note: *der Vorhang* (masc.) and *die Gardine* (fem.) are synonyms meaning *curtain*):
   The curtain rises. It rises. → ~~Der Vorhang~~ *Die Gardine* geht hoch. ~~Er~~ *Sie* geht hoch.

Phrase Addition is applied to all 12,000 ContraPro examples. Depending on suitable conditions, the second and third attack are applied to 3,838 and 1,531 examples, respectively. The Appendix shows results where we vary punctuation and use different added and possessive noun phrases.

5.2.1 Phrase Addition

This attack modifies the previous sentence by appending phrases such as "*...but he wasn't sure*" and also prepending phrases such as "*it is true*:...". A range of other simple phrases can be used, which we leave out for simplicity. In general, all phrases we tried provided lower scores. These attacks introduce a human entity, a pleonastic or an event reference *it* (e.g. *"it is true"*) which are all not plausible antecedents for the anaphoric *it*. We present results for appending "it is true" in Figure 1. Results with using different phrases are presented in the Appendix. In all cases, we prepend or append the same phrase to all ContraPro examples.

5.2.2 Possessive Extension

This attack introduces a new human entity by extending the original antecedent A with a possessive noun phrase e.g., "*the woman's A*". Only two-thirds of the 12,000 ContraPro sentences are linked to an antecedent phrase. Grammar and misannotated antecedents exclude half of the remaining phrases. We put POS-tag constraints on the antecedent phrases before extending them. This reduces our subset to 3,838 modified examples. Our possessive extensions can be humans (*the woman's*), organisations (*the company's*) and names (*Maria's*).

5.2.3 Synonym Replacement

This attack modifies the original German antecedent by replacing it with a German synonym of a different gender. For this we first identify the English antecedent and its most frequent synset in WordNet (Miller, 1995). We obtain a German synonym by mapping this WordNet synsets to GermaNet (Hamp and Feldweg, 1997) synsets. Finally, we modify the correct German pronoun translation to correspond to the gender of the antecedent synonym.

Approximately one quarter of the nouns in our ContraPro examples are found in GermaNet. In 1,531 cases, a synonym of different gender could be identified. Scoring well on the Synonym Replacement attack cannot be done without understanding the pronoun/noun relationship. This attack gets to the core of whether NMT uses CR heuristics instead.

We evaluate a random sample of 100 auto-modified examples as a quality control metric. We note 11 issues with semantically-inappropriate synonyms. Overall, in 14 out of 100 cases, the model switches from correct to incorrect predictions because of synonym-replacement. Only 4 out of these 14 cases come from the questionable synonyms, showing that the drop in ContraPro scores is meaningful.

5.3 Adversarial Attack Results

Our model scores 75.4% on the original ContraPro. This is a very strong result compared to previous work (Müller et al., 2018), largely owing to our model being trained on OpenSubtitles,

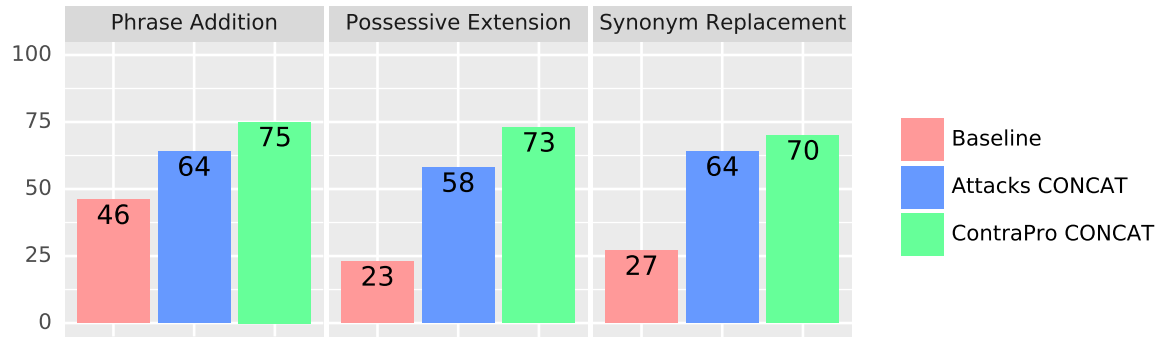

Figure 1: Results with the sentence-level Baseline and CONCAT on ContraPro and three adversarial attacks. The adversarial attacks modify the context, therefore the Baseline model's results on the attacks are unchanged and we omit them. **Phrase**: prepending "it is true: ...". **Possessive**: replacing original antecedent A with "Maria's A". **Synonym**: replacing the original antecedent with different-gender synonyms. Results for Phrase Addition are computed based on all 12,000 ContraPro examples, while for Possessive Extension and Synonym Replacement we only use the suitable subsets of 3,838 and 1,531 ContraPro examples, respectively.

the same domain as the ContraPro examples. The model scores 72.9% and 69.8% on the ContraPro subsets for the Possessive Extension and Synonym Replacement attacks, respectively. The straightforward adversarial modifications we make drop the ContraPro scores by over 10%, as shown by Figure 1. We analyze examples that are scored incorrectly. Some of the attacks introduce an entity that can in principle be referenced by *it*, like extending the antecedent with "*the company's*". In these cases, the new entity's influence on the model is expected, although ideally, the prediction should not change. More surprisingly, attacks that introduce a human entity drop the scores as well. The two largest examples are appending "*...but he wasn't sure*" and extending the original antecedent with *Maria's*. Our synonym replacement leads to a 6% drop in scores.

Intuitively, the adversarial attacks should not contribute to large drops in scores which is contrary to the empirical evidence. Nevertheless, no attack reduces the model's scores close to the original sentence-level baseline. Thus, we conclude that the concatenation model handles CR, but likely with brittle heuristics. Although the results expose potential issues with the model, it is still difficult to pinpoint the specific problems. This reveals a larger issue with pronoun translation evaluation that cannot be addressed with simple adversarial attacks on existing general-purpose challenge sets. We propose 😸, a more systematic approach that targets each of the previously outlined CR pipeline steps with data synthetically generated from corresponding templates.

6 Templates

Automatic adversarial attacks offer less freedom than templates as many systematic modifications cannot be applied to the average sentence. Thus, our 😸 templates are based on the hypothetical coreference pipeline in Section 3 that target each of the three steps: i) Markable Detection, ii) Coreference Resolution and iii) Language Translation. Our minimalistic templates draw entities from sets of 25 animals, 20 human professions (McCoy et al., 2019), 15 foods, and 5 drinks, along with associated verbs and attributes. We use these sets to fill slots in our templates. Animals and foods are natural choices for subject and object slots referenced by *it*. Restricting our sets to interrelated concepts with generically applicable verbs—all animals eat and drink—ensures semantic plausibility. Other object sets, such as buildings, had more semantic implausibility issues and were not included in the final corpus.

Template Target	Example
Priors	
Grammatical Role	The ***cat*** ate the ***egg***. It (😺/🥚) was big.
Order	I stood in front of the ***cat*** and the ***dog***. It (😺/🐶) was big.
Verb	Wow! She unlocked it.
Markable Detection	
Filter Humans	The ***cat*** and the *actress* were happy. However it (😺) was happier.
Coreference Resolution	
Lexical Overlap	The ***cat*** ate the apple and the *owl* drank the water. It (😺) ate the apple quickly.
World Knowledge	The ***cat*** ate the *cookie*. It (😺) was hungry.
Pleonastic it	The *cat* ate the *sausage*. It was raining.
Event Reference	The *cat* ate the *carrot*. It came as a surprise.
Language Translation	
Antecedent Gender	I saw a ***cat***. It(😺) was big. → Ich habe eine Katze gesehen. Sie (😺) war groß.

Table 2: Template examples targeting different CR steps and substeps. For German, we create three versions with *er*, *sie*, or *es* as different translations of *it*.

6.1 Template Generation

Our templates consist of a *previous sentence* that introduces at least one entity and a *main sentence* containing the pronoun *it*. We use contrastive evaluation to judge anaphoric pronoun translation accuracy for each template; we create three translated versions for each German gender corresponding to an English sentence, e.g. *"The cat ate the egg. It rained."* and the corresponding *"Die Katze hat das Ei gegessen.* Er/Sie/Es *regnete"*. To fill a template, we only draw pairs of entities with two different genders, i.e. for animal a and food f: $\text{gender}(a) \neq \text{gender}(f)$. This way we can determine whether the model has picked the right antecedent. We refer to "the model picking an antecedent" as the model scoring the target sentence containing the German third person pronoun with the antecedent's gender higher than the provided alternatives.

First, we create templates that analyze priors of the model for choosing a pronoun when no correct translation is obvious. Then, we create templates with correct translations, guided by the three broad coreference steps. Table 2 provides examples for our templates and the results are shown in Figure 2. Template details—entity sets, statistics, etc.—are provided in the Appendix.

6.1.1 Priors

Prior templates do not have a correct answer, but help to understand the model's biases. We expose three priors with our templates: i) grammatical roles prior (e.g. subject) ii) position prior (e.g. first antecedent) and iii) a general prior if no antecedent and only a verb is present.

For i), we create a Grammatical Role template where both subject and object are valid antecedents. We find that in 72.3% of the template instances, the model chooses the object as the antecedent.

For ii), we create a Position template where two objects are enumerated (see Table 2). We create an additional example where the entities order is reversed and test if there are priors for specific nouns or alternatively positions in the sentence.

The model shows a strong prior for neuter by predicting *es* in most cases, even if the two entities are masculine and feminine.

For iii), we create a Verb template, expecting that certain transitive verbs trigger certain object gender choices. We use 100 frequent transitive verbs and create sentences such as the example in Table 2. As expected, *it* is translated to the neuter *es* most of the time, with notable exceptions where the verb is strongly associated with a single noun, e.g. *"Sie hat sie entriegelt"* is scored higher for *"She unlocked it"*. We presume that the reason for this is that *to unlock a door* is very common and door (*Tür*) is feminine in German.

6.1.2 Markable Detection with a Humanness Filter

Before doing the actual CR, the model needs to identify all possible entities that *it* can refer to. We construct a template that contains a human and animal which are in principle plausible antecedents, if not for the condition that *it* does not refer to people. For instance, the model should always choose *cat* in *"The* actress *and the* cat *were hungry. However* it *was hungrier."*. We find that the model instead falls back to translating *it* to the neuter *es* in all cases.

6.1.3 Coreference Resolution

Having determined all possible antecedents, the model has to choose the correct one, relying on semantics, syntax, and discourse. The pronoun *it* can in principle be used as an *anaphoric* (referring to entities), *event reference* or *pleonastic* pronoun (Loáiciga et al., 2017). For the anaphoric *it*, we identify two major ways of identifying the antecedent: lexical overlap and world knowledge. Our templates for these categories are meant to be simple and solvable.

Overlap: Broadly speaking the subject, verb, or object can overlap from the previous sentence to the main sentence, as well as combinations of them. This gives us five templates: i) subject-overlap ii) verb-overlap iii) object-overlap iv) subject-verb-overlap and v) object-verb-overlap. We always use the same template for the context sentence. e.g. *"The* ***cat*** *ate the apple and the* ***owl*** *drank the water."*. For the object-verb-overlap we would then create the main sentence *"It ate the apple quickly."* and expect the model to choose *cat* as antecedent. To keep our overlap templates order-agnostic, we vary the order in the previous sentence by also creating *"The* ***owl*** *drank the water and the* ***cat*** *ate the apple."* However our results in 6.2 show that the model's predictions are almost completely random and are influenced by position priors, e.g., the first mentioned subject, or a prior for the neuter *es* when it needs to decide between the two subjects.

World Knowledge: CR has been traditionally seen as challenging as it requires world knowledge. Our templates test simple forms of world knowledge by using attributes that either apply to animal or food entities, such as *cooked* for food or *hungry* for animals. We then evaluate whether the model chooses e.g. *cat* in *"The* ***cat*** *ate the cookie. It was hungry."* As discussed later, the model occasionally predicts answers that require world knowledge, but most predictions are guided by a prior for choosing the neuter *es* or a prior for the subject.

Pleonastic and Event Templates: For the other two ways of using *it*, event reference and pleonastic-it, we again create a default previous sentence (*"The* ***cat*** *ate the* apple.*")*. For the main sentence, we used four typical pleonastic and event reference phrases such as *"It is a shame"* and *"It came as a surprise"*. We expect the model to correctly choose the neuter *es* as a translation every time and the strong prior for the neuter gender causes the model to do so nearly perfectly.

6.1.4 Translation to German

After CR, the decoder has to translate from English to German. In our contrastive scoring approach the translation of the English antecedent to German is already given. However the decoder is still required to know the gender of the German noun to select between *er, sie* or, *es*. We test this with a list of concrete nouns selected from Brysbaert et al. (2014), which we filter for nouns that occur more than 30 times in the training data. We are left with 2051 nouns which are plugged into the *"I saw a N. It was {big, small}."* template.

6.2 Results

We find that the model performs poorly when actual CR is required. It frequently falls back to choosing the neuter *es* or preferring a position (e.g. first of two entities) for determining the gender. For *Markable Detection* the model always predicts the neuter *es* regardless of the actual genders of the entities.

In the Overlap template, we find that the model fails to recognize the overlap and instead, has a general preference for one of the two clauses. For instance in the case of verb-overlap, the model had a solid accuracy of 64.1% if the verb overlapped from the first clause (*"The cat ate*

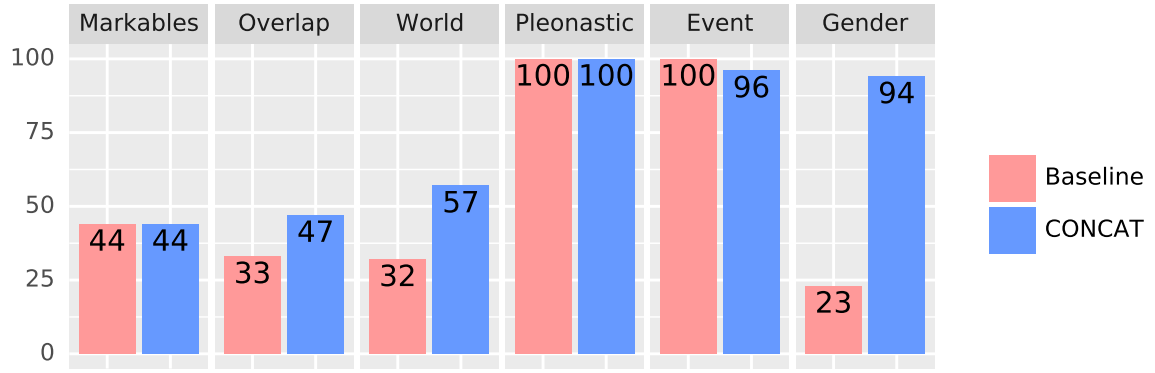

Figure 2: Results comparing the sentence-level baseline to CONCAT on ContraCAT. Pronoun translation pertaining to World Knowledge and language-specific Gender Knowledge benefits the most from additional context.

and the dog drank. It ate a lot.") but a weak accuracy of 39.0% when the verb overlapped from the second clause (*"The cat ate and the dog drank. It drank a lot."*) The overall accuracy for the overlap templates is 47.2%, with little variation across the types of overlap. Adding more overlap, e.g., by overlapping both the verb and object (*"It ate the apple happily"*), yields no improvement. Overall, the model pays very little attention to overlaps when resolving pronouns.

We also see weak performance for world knowledge. An accuracy of 55.7% is slightly above the heuristic of randomly choosing an entity (= 50.0%). With a strong bias for the neuter *es*, the model has a high accuracy of 96.2% for event reference and pleonastic templates, where *es* is always the correct answer. Based on the strong performance on the Gender template in 6.1.4, we conclude the model consistently memorized the gender of concrete nouns. Hence, CR mistakes stem from Step 1 or Step 2, suggesting that the model failed to learn proper CR.

7 Augmentation

We present an approach for augmenting the training data. While challenging for NLP, we focus on a narrow problem which lends itself to easier data manipulation. Our previous analyses show that our model is capable of modeling the gender of nouns. However, they also show a strong prior to translate *it* to *es* and very little CR capability. Our goal with the augmentation is to break off the strong prior and test if this can give rise to better CR in the model.

We attempt to do this by augmenting our training data and call it Antecedent-free augmentation (AFA). We identify candidates for augmentation as sentences where a coreferential *it* refers to an antecedent not present in the current or previous sentence (e.g., *I told you before. <SEP> It is red. → Ich habe dir schonmal gesagt. <SEP> Es ist rot.*). We create augmentations by adding two new training examples where the gender of the German translation of "it" is modified (e.g., the two new targets are "*Ich habe dir schonmal gesagt. <SEP> Er ist rot.*" and "*Ich habe dir schonmal gesagt. <SEP> Sie ist rot.*"). The source side remains the same. An additional example is shown in Table 3. Antecedents and coreferential pronouns are identified using a CR tool (Clark and Manning, 2016a; Clark and Manning, 2016b). We fine-tune our already trained concatenation model on a dataset consisting of the candidates and the augmented samples. As a baseline, we fine-tune on the candidates only so as to confidently say that any potential improvements come from the augmentations.

7.1 Results

7.1.1 Adversarial Attacks

AFA provides large improvements, scoring 85.3% on ContraPro. Results are shown in Figure 3. The AFA baseline (fine-tuning on the augmentation candidates only) improves by 1.94%,

Antecedent-free augmentation	
Source	You let me worry about that. <SEP> How much you take for it?
Reference	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet er?
Augmentation 1	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet sie?
Augmentation 2	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet es?

Table 3: Examples of training data augmentations. The source side of the augmented examples remains the same.

presumably because many candidates consist of coreference chains of "it" and the model learns they are important for coreferential pronouns. However, the improvement is small compared to AFA.

Results on ContraPro for each gender (see Appendix) show that performance on *er* and *sie* is substantially increased, suggesting that the augmentation successfully removes the strong bias towards *es*. Templates provide further evidence about this. Although, the adversarial attacks lower AFA scores, in contrast to CONCAT, the model is more robust and the performance degradation is substantially lower (except on the synonym attack). We experimented with different learning rates during fine-tuning and present results with the LR that obtained the best baseline ContraPro score. Detailed scores in the Appendix show how LR can balance the scores across the three different genders. Furthermore, CONCAT and AFA obtain 31.5 and 32.2 BLEU on ContraPro, respectively, showing that this fine-tuning procedure, which is tailored to pronoun translation, does not lead to any degradation in general translation quality.

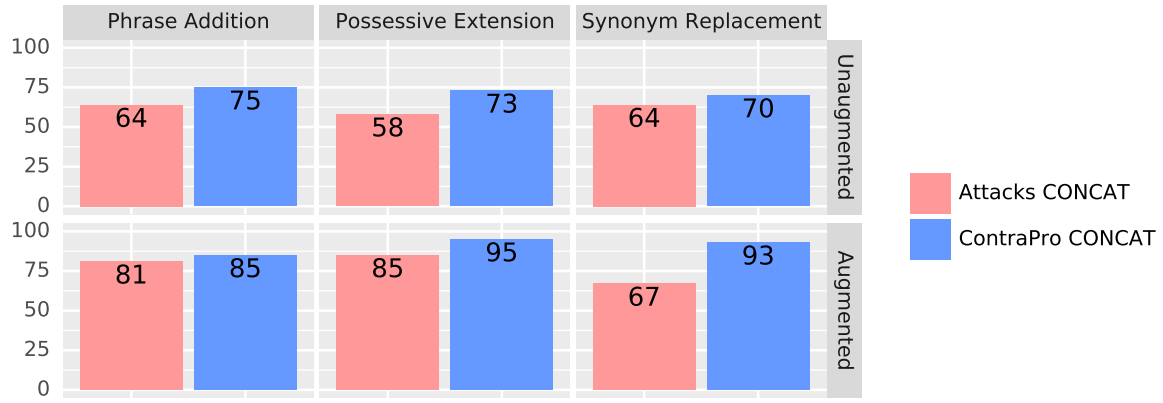

Figure 3: Results comparing unaugmented and augmented CONCAT on ContraPro and same 3 attacks as in Figure 1. Results with non-augmented CONCAT are the same as Figure 1.

7.1.2 Templates

From the prior templates, we observe that the prior over gender pronouns is more evenly spread and not concentrated on *es*. This also provides for a more even distribution on the Position and Role Prior template. The results on the prior templates are presented in the Appendix. The augmented model is also substantially better on markable detection, improving by 27.6%. Results for templates are presented in Figure 4.

No improvements are observed on the World Knowledge template. Pleonastic cases are still reasonably handled, although not perfectly as with CONCAT. The Event template identifies a systematic issue with our augmentation. We presume this is as a result of the CR tool marking cases where *it* refers to events. We do not apply any filtering and augment these cases as well, thus create wrong examples (an event reference *it* cannot be translated to *er* or *sie*). As a result, the scores are significantly lower compared to CONCAT. We note that this issue with our model is not visible on ContraPro and the adversarial attacks results. In contrast, the Event template

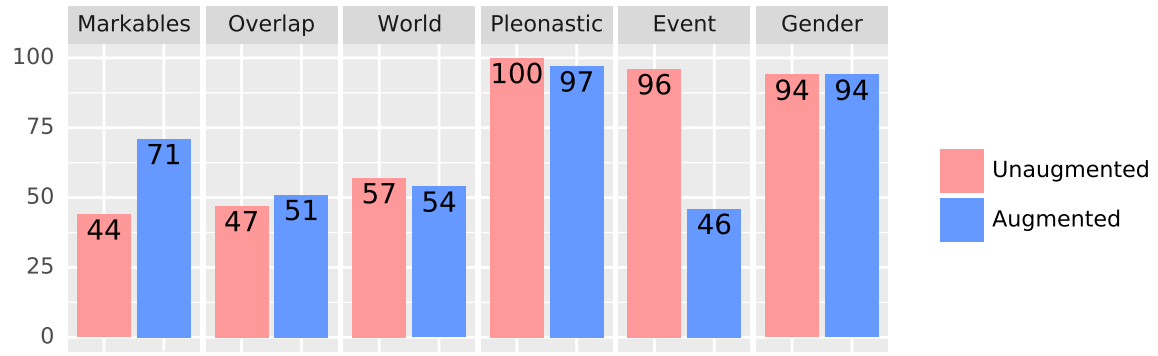

Figure 4: ContraCAT results with unaugmented and augmented CONCAT. We speculate that readjusting the prior over genders in augmented CONCAT explains the improvements on Markable and Overlap.

easily identifies this problem.

AFA performs on par with the unaugmented baseline on the Gender template. However, despite increasing by 3.8%, results on Overlap are still underwhelming. Our analysis shows that augmentation helps in changing the prior. We believe this provides for improved CR heuristics which in turn provide for an improvement in coreferential pronoun translation. Nevertheless, the Overlap template shows that augmented models still do not solve CR in a fundamental way.

8 Conclusion

In this work, we study how and to what extent CR is handled in context-aware NMT. We show that standard challenge sets can easily be manipulated with adversarial attacks that cause dramatic drops in performance, suggesting that NMT uses a set of heuristics to solve the complex task of CR. Attempting to diagnose the underlying reasons for these results, we propose targeted templates which systematically test the different aspects necessary for CR. This analysis shows that while some type of CR such as pleonastic and event CR are handled well, NMT does not solve the task in an abstract sense. We also propose a data augmentation approach which substantially improves performance on some metrics, but it does not change the general conclusions we infer from the templates. Future work should be evaluated on our adversarial attacks and ContraCAT, which we publicly release, to realistically estimate the ability of NMT to robustly do CR.

Acknowledgments

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement № 640550). This work was also supported by DFG (grant FR 2829/4-1). We thank Alexandra Chronopoulou for the valuable comments and helpful feedback.

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, June. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods*, 46:904–911.

Kevin Clark and Christopher D. Manning. 2016a. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas, November. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, August. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

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, May. 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, October-November. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Christian Hardmeier. 2012. Discourse in Statistical Machine Translation. A Survey and a Case Study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *ArXiv e-prints*, December.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China, November. Association for Computational Linguistics.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. 2006. Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding. In *Final Report of the 2006 JHU Summer Workshop*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the 15th conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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, September. Association for Computational Linguistics.

António Lopes, M Amin Farajian, Rachel Bawden, Michael Zhang, and André Martins. 2020. Document-level Neural MT: A Systematic Comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective Attention for Context-aware Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954. 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, September. Association for Computational Linguistics.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

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 *WMT 2018*, Brussels, Belgium. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July. 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.

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, April. 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, July. Association for Computational Linguistics.

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, October. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2019. Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 140–150, Dublin, Ireland, August. European Association for Machine Translation.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

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*, pages 6000–6010.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

Xiaosen Wang, Hao Jin, and Kun He. 2019. Natural Language Adversarial Attacks and Defenses in Word Level. *arXiv preprint arXiv:1909.06723*.

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*.

A Preprocessing Details and Model Hyper-Parameters

We use OpenSubtitles2018[2] (Lison and Tiedemann, 2016) as training data. We tokenize the dataset using the Moses scripts[3] (Koehn et al., 2006). We BPE-split the data by jointly computing them on English and German using 32K merge operations. We remove all samples where the main sentence exceeds 100 tokens on the BPE-level or the concatenated sample contains more than 200 tokens. ContraPro is built using OpenSubtitles and contains samples from it. As a result, we remove the entire documents from which the ContraPro samples originate from in order to remove any exact duplicates from ContraPro or similar contexts which may lead to unfair advantages for the models. This still leaves some exact duplicates between our training data and ContraPro which we also remove. The model is finally trained on ≈16.7M samples.

We train the transformer models with a batch size of 4096. We use an initial learning rate of 10^{-4} and we lower it by a factor of 0.7 if there are no improvements on the validation perplexity for 8 checkpoints. We save a checkpoint every 4000 updates.

The transformer models we use are a 6 layer encoder/decoder with 8 attention heads. The model size is 512 and the size of the feed-forward layers is 2048. We tie the source, target and output embeddings. We use label smoothing with 0.1 and dropout in the transformer of 0.1. Models are trained on 2 GTX 1080 GPUs with 8GB RAM. The final model is an average of the 8 best checkpoints based on validation perplexity. The models we train are implemented in Sockeye (Hieber et al., 2017).

B Complete List of Automatic Attacks on ContraPro

B.1 Adding Phrases

As a reference to the scores shown in Table 4, our model has a score of 0.754 on unmodified ContraPro. C is the original context sentence from ContraPro.

B.2 Possessive Extension

These were applied to 3,838 ContraPro examples. As a reference to scores shown in Table 5, our model has a score of 72.9% on the unmodified subset of ContraPro. A refers to the original antecedent noun phrase.

[2] `http://opus.nlpl.eu/OpenSubtitles-v2018.php`
[3] `https://github.com/moses-smt/mosesdecoder`

Modification	ContraPro Score
he/she said: "*C*"	66.5 / 66.3
it is true:"*C* / it is true: *C*	55.2 / 63.5
C and it is true. / *C*. it is true. / *C* and that is true.	65.1 / 57.9 / 70.4
C but he wasn't sure. / *C*. but he wasn't sure.	69.0 / 65.8
C but that's not the point. / *C*. but that's not the point.	70.7 / 67.5
C but there is a catch. /*C*. but there is a catch.	67.4 / 64.6
C but why. / *C*. but why.	74.0 / 71.5

Table 4: Scores for each *Adding-Phrase*-modification. Slightly altered modifications are indicated with "/"

Modification	ContraPro Score
... the woman's *A* ...	63.4
... the man's *A* ...	66.5
... my mother's *A* ...	70.9
... my father's *A* ...	72.2
... the dog's *A* ...	66.8
... the cat's *A* ...	67.4
... the doctor's (vom Arzt/von der Ärztin) *A* ...	66.7 / 66.4
... *A* of my best friend's mother ...	60.0
... the government's *A* ...	68.5
... the company's *A* ...	63.0
... Maria's *A* ...	58.3
... Lisa's *A* ...	60.3
... Bolsena's *A* ...	60.3
... Peter's *A* ...	59.0
... Robert's *A* ...	60.5
... David's *A* ...	60.6

Table 5: Scores for each *Possessive-Extension*-modification. For German we append the possessive noun phrase with "von"(=of).

B.3 Synonym replacement

These were applied to 1,531 ContraPro examples. Our model has a score of 69.8% on the unmodified subset of ContraPro. When replacing with different-gender synonyms we drop to a score of 64.1%.

C Template Generation

C.1 Vocabulary

Our templates draw from the sets of entities shown in Table 6 and Table 7. The translations in German are shown in brackets. We note that all entities appear in the training dataset we use to train our models. The least frequent entity ("kangaroo") appears 134 times.

We also use four event- and pleonastic-it phrases which are used as the main sentence in the templates and referred to later.

Event: It came as a surprise (Es kam überraschend), It actually happened (Es ist tatsächlich passiert), It resulted in chaos (Es führte zu Chaos), It was a funny situation (Es war eine lustige Situation)

Pleonastic: It was raining (Es regnete), It is a shame (Es ist eine Schande), It seemed this was unnecessary (Es schien, dass dies unnötig war), It is hard to believe this is true (Es ist schwer zu glauben , dass das wahr ist)

ANIMALS		PROFESSIONS
dog (Hund)	giraffe (Giraffe)	professor (Professor(in))
wolf (Wolf)	mouse (Maus)	student (Student(in))
bear (Bär)	duck (Ente)	judge (Richter(in))
tiger (Tiger)	turtle (Schildkröte)	secretary (Sekretär(in))
lion (Löwe)	owl (Eule)	doctor (Arzt/Ärztin)
rabbit (Hase)	dove (Taube)	lawyer (Anwalt/Anwältin)
monkey (Affe)	goat (Ziege)	scientist (Wissenschaftler(in))
eagle (Adler)	sheep (Schaf)	manager (Manager(in))
frog (Frosch)	squirrel (Eichhörnchen)	artist (Künstler(in))
cat (Katze)	horse (Pferd)	actor (Schauspieler)
cow (Kuh)	pig (Schwein)	actress (Schauspielerin)
zebra (Zebra)	kangaroo (Känguru)	
deer (Reh)		

Table 6: Vocabulary of entities used in templates.

FOOD		DRINKS	
cookie (Keks)	cake (Kuchen)	tea (Tee)	juice (Saft)
carrot (Karotte)	hot dog (Hotdog)	milk (Milch)	lemonade (Limonade)
cheese (Käse)	apple (Apfel)	water (Wasser)	
nut (Nuss)	fruit (Frucht)		
sausage (Wurst)	pizza (Pizza)		
bread (Brot)	egg (Ei)		
meat (Fleisch)	ice cream (Eis)		
steak (Steak)			

Table 7: Vocabulary of entites used in templates.

C.2 Template Statistics

For each template, we report the number of lines it contains in Table 8.

Template	Number of lines
Grammatical Role Prior	1000
Position Prior	828
Verb Prior	600
Markable Detection (animacy)	2560
Verb Overlap	2240
Object Overlap	5376
Subject Overlap	4992
Object-Verb Overlap	5376
Subject-Verb Overlap	4992
World Knowledge	2500
Event	1500
Pleonastic	1500
Gender	4102

Table 8: Number of test sentences for each template.

C.3 Template Definitions

The template definitions are shown in Table 9. We refer to animals with A, professions as P, food as F, drinks as D. When creating a concrete animal, food or drink X_i, we use the definite

article "the" ("der/die/das" in German). On the German side, we underline the options that we give the model for the three German genders.

Template	English definition	German definition
Grammatical Role Prior	A ate F. It was {big, small, large, tiny}.	A hat F gegessen. Er/Sie/Es war {groß, klein, riesig, winzig}.
Position Prior	I stood in front of A_i and A_j. It was {big, small, large, tiny}.	Ich stand vor A_i und A_j. Er/Sie/Es war {groß, klein, riesig, winzig}.
Verb Prior	Wow! I/You/He/She/We/They $V_{past+transitive}$ it.	Wow! Ich/Du/Er/Sie/Wir/Sie haben er/sie/es $V_{past+transitive}$.
Markable Detection (filter humans)*	A and P were {hungry, tired, happy, nice}. However it was {hungrier, more tired, happier, nicer}.	A und P waren {hungrig, müde, glücklich, nett}. Aber er/sie/es war {hungriger, müder, glücklicher, netter}.
Verb Overlap*	A_i {ate, drank} and A_j {ate, drank}. It {ate, drank} {a lot, quickly, slowly happily}.	A_i hat {gegessen, getrunken} und A_j hat {gegessen, getrunken}. Er/Sie/Es hat {viel, schnell, langsam, fröhlich} {gegessen, getrunken}.
Object Overlap*	A_i ate F and A_j drank D. It liked {F, D}.	A_i hat F gegessen und A_j hat D getrunken. Er/Sie/Es mochte {F, D}.
Subject Overlap*	A_i ate F and A_j drank D. {A_i, A_j} liked it.	A_i hat F gegessen und A_j hat D getrunken. {A_i, A_j} mochte ihn/sie/es.
Object-Verb Overlap*	A_i ate F and A_j drank D. It {ate F, drank D} quickly.	A_i hat F gegessen und A_j hat D getrunken. Er/Sie/Es hat {F schnell gegessen, D schnell getrunken}.
Subject-Verb Overlap*	A_i ate F and A_j drank D. {A_i ate, A_j drank} it quickly.	A_i hat F gegessen und A_j hat D getrunken. {A_i, A_j} hat ihn/sie/es schnell {gegessen, getrunken}.
World Knowledge	A ate F. It {was hungry, was looking around, was running around, was tired, was happy} / {had a sweet/bitter/sour taste, was cooked, had gone bad}.	A hat F gegessen. Er/Sie/Es {war hungrig, schaute sich um, rannte herum, war müde, war glücklich} / {hatte einen süßen/bitteren/sauren Geschmack, war gekocht, war schlecht geworden}.
Event	A ate F. EVENT-PHRASE	A hat F gegessen. EVENT-PHRASE.
Pleonastic	A ate F. PLEONASTIC-PHRASE	A hat F gegessen. PLEONASTIC-PHRASE.
Gender	I saw a $N_{concrete}$. It was {big, small}.	Ich sah ein/eine/einen $N_{concrete}$. Er/Sie/Es war {groß, klein}.

Table 9: Template definitions. * We switch the position (first or second) of the two involved entities E_i and E_j.

C.4 Prior Results

For the templates that do have a correct answer, we show results in the main paper. In Table 10, Table 11 and Table 12 we show the results on the grammatical, position and verb prior templates.

Model	subject	object
CONCAT	20.7%	72.3%
AFA	52.2%	47.8%

Table 10: Grammatical Role template for testing prior of choosing subject or object as antecedent to translate *it*. If numbers do not add up to 100%, it is because the model chose neither the subject nor object. This is usually the neuter *es*.

Model	first	second	same antecedent
CONCAT	0.0%	3.1%	60.8%
AFA	0.2%	13.0%	74.9%

Table 11: Position template for testing prior for first or second enumerated object as antecedent to translate *it*. If numbers do not add up to 100%, it is because the model chose neither the first nor second object. This is usually the neuter *es*.

Model	masculine	feminine	neuter
CONCAT	5.7%	2.8%	91.5%
AFA	43.7%	27.5%	28.8%

Table 12: Verb template for testing prior for the three genders, only conditioned on a transitive verb.

D Augmentation

D.1 Details

For all augmentations we use Spacy's dependency parser[4] in order to determine the case of the pronoun. This is necessary because the feminine ("sie") and neuter ("es") pronoun are the same in nominative and accusative, but the masculine is not ("er" and "ihn"). We fine-tuned on 207K for the antecedent-free augmentations.

D.2 Fine-Tuning Learning Rate Analysis

We conducted 3 different fine-tuning experiments where we varied the learning rate. We used a learning rate of $2 * 10^{-6}$, $2 * 10^{-7}$ and $2 * 10^{-8}$. The initial concatenation model was trained with an initial LR of $2 * 10^{-4}$ and when it converged, the learning rate was $7.82 * 10^{-8}$. Results are presented in Table 13. As before, we average 8 checkpoints before evaluating our models.

	total	er	sie	es
CONCAT	75.4	64.0	66.8	95.3
AFA lr=10^{-6}	78.4	81.0	81.9	72.5
AFA lr=10^{-7}	85.3	88.2	90.6	77.2
AFA lr=10^{-8}	81.3	73.9	77.3	92.7

Table 13: Challenge set performance for each pronoun.

[4] https://spacy.io/usage/linguistic-features#dependency-parse

As the goal with the augmentations is to remove the strong bias towards neuter, we only evaluate the different LR models on the "er", "sie" and "es" accuracy on ContraPro. Performance on "er" and "sie" improves in all experiments, but it improves by far the most using a LR of $2 * 10^{-7}$. Performance on "es" gets worse as the LR increases. However, very low LR also does not provide for large improvements on "er" and "sie". We show that the LR is an important hyper-parameter in order to balance the performance on all pronouns. Admittedly, one may opt for a lower learning rate because, as the training data shows, "it" tends to be translated to "es", so it is undesirable to significantly drop performance on "es" because in practice these errors will be more visible.