Chapter 1: See Other File

Chapter 2: See Other File

Chapter 3: See Other File

Chapter 4: Automatic Data Generation without a Source

Chapter **??** introduces the idea of automation for **generating** data. However, in that project there was a source of **found** data. How can one automate data **generation** without an existing source? This chapter uses an **expert** to design a series of rules to automatically **generate** a dataset. The limitations of the automatic data will motivate using crowd-sourcing in Chapter **??**. The merit of the **expert** in the process will motivate using them directly in Chapter **??**.

4.1 Evaluating Data

Genuinely varied, realistic data is necessary to create models that are robust to minor variations. However, equally robust evaluation methodologies are important in ascertaining the quality of the data. Current methods focus on quantitative assessments that may inadvertently assess the **annotation**, but not the **generation**, quality of a dataset. Since most datasets are evaluated on the same types of data—SQuAD test data is comparable to the training data—the linguistic variation of a dataset is not readily captured by standard quantitative metrics like accuracy or F_1. Furthermore, a model that has memorized several key answers upon which it is then tested is not necessarily *learning*; raw analysis of data overlap confirms this risk (Lewis et al., 2020). Datasets meant to effectively and robustly evaluate trained datasets can determine how much of a problem this poses *ex-post-facto*.

As one solution to this limitation, Checklist (Ribeiro et al., 2020) created a task-agnostic methodology for testing NLP models. We extend this work to a specific task: testing coreference (Soon et al., 2001) in machine translation. There does not exist a dataset that can serve as a source, unlike our past automation work (Chapter **??**). The dataset we create is designed by **experts**: specifically native German and native English speakers, and scaled through **automation**. While a similar dataset of the same size could be created without knowledge of either language, the templates used as test data would prove be nonsensical or unnatural.

4.2 Meaningful Model Evaluation in Machine Translation

Machine translation is a classic and complex NLP task that requires diverse linguistic knowledge and data in multiple languages. Classic datasets were often gathered through extensive collaboration with **experts**. However, recent ones are often created through **crowd-sourcing** or **automatic** methods. Therefore, this is an area well-suited to our evaluation techniques.

We focus on German-English coreference resolution as a representative task. 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). A German pronoun can have three genders, determined by its antecedent: masculine (*er*), feminine (*sie*) and neuter (*es*). The nuance of this work requires native knowledge of both English and German.

Accuracy in machine translation is at an all-time high with the rise of neural architectures (Wu et al., 2016) but this metric alone is insufficient for evaluation. Previous work (Hardmeier and Federico, 2010; Miculicich Werlen and Popescu-Belis, 2017; Müller et al., 2018) proposes evaluation methods for specifically pronoun translation. Context-aware neural machine translation (NMT) models are capable of using discourse-level information and are prime candidates for this evaluation. We ask:

> 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 propose extending ContraPro (Müller et al., 2018)—a contrastive challenge set for automatic English→German pronoun translation evaluation[1]—by making small adversarial changes in the contextual sentences.[2]

Our adversarial attacks on ContraPro show 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. In our case, trivial changes in language cause incorrect predictions, but both the changes and the prediction would not be noticed by somebody without a mastery of German. NMT uses brittle heuristics to solve CR if trivial changes in pronouns and nouns fool a coreference corpus like ContraPro. However, this will not identify *which* heuristics these are.

For this reason, we propose a new dataset, created from templates (Section 4.6.3.2), to systematically evaluate which heuristics are being used in coreferential pronoun translation. Inspired by previous work on CR (Raghunathan et al., 2010; Lee et al., 2011), we create templates tailored to evaluating the specific steps of an idealized CR pipeline. We call this collection ContraCAT, **C**ontrastive **C**oreference **A**nalytical **T**emplates. The construction of templates is controlled, enabling us to easily create large number of coherent test examples and provide unambiguous conclusions about the CR capabilities of NMT. While this methodology depends on

[1] ContraPro is described in Section 4.6.1.

[2] Equal effort between Denis Peskov, Benno Krojer, Dario Stojanovski, and supervised by Alex Fraser. 2020. In International Conference on Computational Linguistics

Peskov is responsible for part of template design, selecting concrete nouns for the templates, paper writing, and the video.

automation, a technique called into question in Chapter **??**, the templates are written in collaborations between a native German speaker and native English speakers. Since automation is subject to quality control issues, this level of expertise is necessary if the adversarial dataset is to be reflective of actual language used by English and German speakers. The procedure used in creating these templates can be adapted to many language pairs with little effort. We also propose a simple data augmentation approach using fine-tuning. This methodology should not change the way CR is being handled by NMT and support the hypothesis that automated data techniques have limited applicability. We release a new dataset, ContraCAT, and the adversarial modifications to ContraPro.

ContraCAT applies only to coreference, but the investigation of heuristics is an important research direction in NLP that can measure the issues noted with **automatic** (Chapter **??**) and **crowd-sourced** (Chapter **??**) datasets. Heuristics are accurate if there are underlying data limitations; this implies that the training data and the evaluation data resemble one another in superficial ones. Therefore, exposing the brittleness in current datasets motivates the need for higher-quality evaluation data—to observe this limitation—and varied training data—to overcome it.

We introduce coreference resolution as a task in Section 4.3, the idealized coreference pipeline in Section 4.3, and the transformer model in Section 4.5. We discuss ContraPro in Section 4.6.1, and explain our proposed templates in Section 4.6.2.

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

Table 4.1: A hypothetical CR pipeline that sequentially resolves and translates a pronoun.

4.3 Why is Coreference Resolution Relevant?

Evaluating discourse phenomena is an important first step in evaluating 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, 2018; Miculicich et al., 2018; Voita et al., 2019; Maruf et al., 2019).

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 will be *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. This is an evaluation metric applicable for multiple forms of **generated** NLP data.

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).

4.4 Do Androids Dream of Coreference Translation Pipelines?

Imagine a hypothetical coreference pipeline that generates a pronoun in a target language, as illustrated in Table 4.1. *First*, markables (entities that can be referred to by pronouns) are tagged in the source sentence (we restrict ourselves to concrete entities as concepts are incompatible with many verbs). 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 addressing 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 CorefAnnotator (Raghunathan et al., 2010; Lee et al., 2011). However, NMT models are unable to decouple the individual steps of this pipeline. We propose to isolate each of these steps through targeted examples.

4.5 Model

We use a transformer model (Background Section **??**) 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.[3]

[3] The training examples for this model are modified by prepending the previous source and target sentence to the main source and target sentence. 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.

4.6 Adversarial Attacks

ContraPro (Müller et al., 2018), a contrastive challenge set (Section 4.2), has limitations; the methodology for creating our own dataset addresses them.

4.6.1 About ContraPro

ContraPro is a contrastive challenge set for English→German pronoun translation evaluation. This dataset is **automatically** generated (Chapter **??**), making it subject to manipulation and preventing it from fully elucidating the limitations of neural coreference resolution. The set consists of English sentences containing an anaphoric pronoun *it* and the corresponding German translations (e.g., "*Give me your hand, ah, it's soft and hot, and it feels pleasant*"→"*Gib deine Hand, ah, sie ist weich und warm, und wohlig fühlt sie sich an.*"). 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 the theoretically inconsequential parts of the context sentence before the occurrence of *it*. Coreference accuracy degrades from this adversarial attack *contrary* to the expectation that a transformer model would discard inconsequential priming.

4.6.2 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 original antecedent with 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 original German antecedent with synonym of different gender:

   The curtain rises. It rises. → ~~Der Vorhang~~ *Die Gardine* geht hoch. ~~Er~~ *Sie* geht hoch.[4]

Phrase Addition can be applied to all 12,000 ContraPro examples. The second and third attack can only be applied to 3,838 and 1,531 examples, due to the required sentence contingencies.

[4] *der Vorhang* (masc.) and *die Gardine* (fem.) are synonyms meaning *curtain*

4.6.2.1 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. All phrases we tried provided lower scores. These attacks either introduce a human entity or an event reference *it* (e.g., *"it is true"*) which are both not plausible antecedents for the anaphoric *it*.

4.6.2.2 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 filters our subset to 3,838 modified examples. Our possessive extensions can be humans (*the woman's*), organisations (*the company's*) and names (*Maria's*).

4.6.2.3 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 Feld-

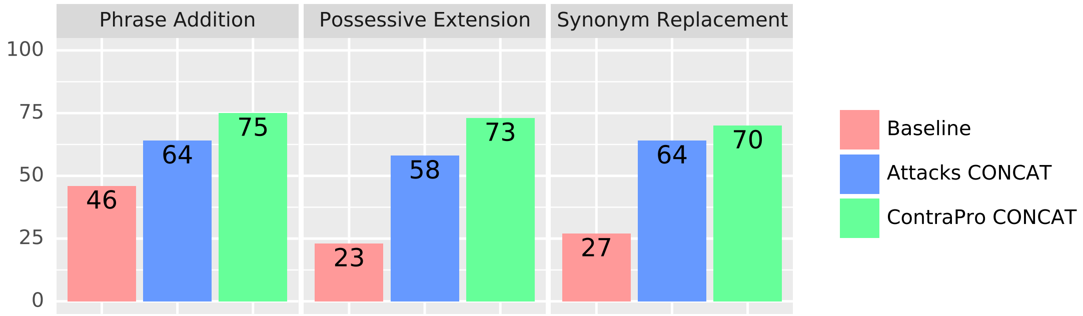

Figure 4.1: Results with the sentence-level Baseline and CONCAT on ContraPro and three adversarial attacks. The adversarial attacks modify the context, therefore the Baseline model's results on the attacks are unchanged and we omit them. **Phrase**: prepending "it is true: . . . ". **Possessive**: replacing 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.

weg, 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 of these cases, a synonym of different gender could be identified. The Synonym Replacement attack gets to the core of whether NMT uses CR heuristics as understanding the pronoun-noun relationship is paramount to predicting the correct pronoun.

4.6.3 Quality Assessment of the Automatic Attacks by an Expert

We evaluate a random sample of 100 auto-modified examples as a quality control metric. There are 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.

4.6.3.1 Evaluating Adversarial Attacks

Intuitively, the adversarial attacks should not contribute to large drops in scores, since no meaningful changes are being made. If the model accuracy drops some, but not all the way to the original sentence-level baseline (Section 4.5), we can conclude that the concatenation model handles CR, but likely with brittle heuristics. If the model accuracy drops all the way to the baseline, then the model is memorizing the inputs. The changes in accuracy suggest issues, but not to ascertain what they are. 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 ContraCAT, a more systematic approach that targets each of the previously outlined CR pipeline steps with data synthetically generated from corresponding templates.

Automatic adversarial attacks offer less freedom than templates as many systematic modifications cannot be applied to the average sentence. Thus, our ContraCAT templates are built on the hypothetical coreference pipeline in Section 4.4

that target each of the three steps: 1) Markable Detection, 2) Coreference Resolution and 3) Language Translation. Our minimalistic templates draw entities from sets of animals, human professions (McCoy et al., 2019), foods, and 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, would cause semantic implausibility with certain verbs.

4.6.3.2 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: gender(a) $\neq$ gender(f). This way we can determine whether the model has picked the right antecedent.

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 4.2 provides examples for our templates.

4.6.3.3 Priors

Our templates that test prior biases do not have a correct answer but reveal the model's biases. We expose three priors with our templates: 1) grammatical roles prior (e.g., subject) 2) position prior (e.g., first antecedent) and 3) a general prior if no antecedent and only a verb is present.

For the first prior, we create a Grammatical Role template where both subject and object are valid antecedents.

For the second prior, we create a Position template where two objects are enumerated as shown in Table 4.2. 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.

For the third prior, we create a Verb template, expecting that certain transitive verbs trigger certain object gender choice. We use 100 frequent transitive verbs and create sentences such as the example in Table 4.2.

4.6.3.4 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 *are hungry. However* it *is hungrier.*".

4.6.3.5 Coreference Resolution

Having determined all possible antecedents, the model chooses 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: subject-overlap, verb-overlap, object-overlap, subject-verb-overlap and 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."*

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."* 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.

4.6.3.6 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. This selects 2051 nouns that are substituted for N in: *"I saw a N. It was {big, small}."*.

4.6.4 Results

The CONCAT model becomes less accurate 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.

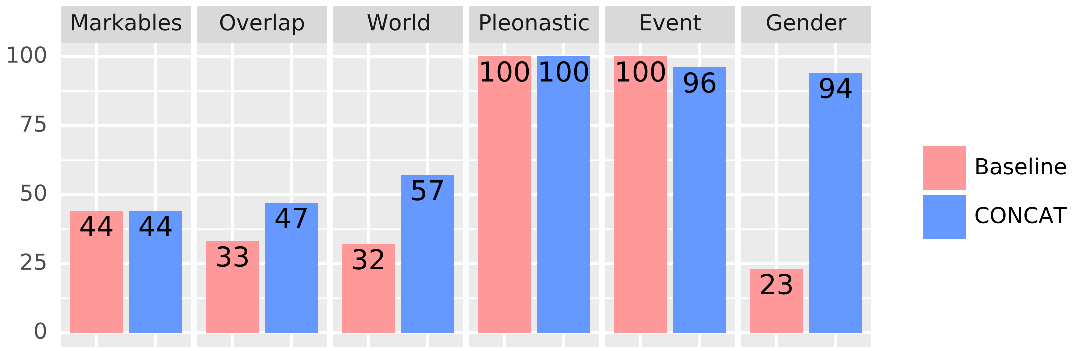

Figure 4.2: Results comparing the sentence-level baseline to CONCAT on ContraCAT. Pronoun translation pertaining to World Knowledge and language-specific Gender Knowledge benefits the most from additional context.

In the **Overlap** template, the model fails to recognize the overlap and has a general preference for one of the two clauses. In the case of verb-overlap, the model has an accuracy of 64.1% if the verb overlapped from the first clause (*"The cat ate and the dog drank. It ate a lot."*), but a low 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 little attention to overlaps when resolving pronouns.

We also see weak accuracy on tests of world knowledge. An accuracy of 55.7% is slightly above the heuristic of randomly choosing an entity (= 50.0%). Due to the 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 high accuracy on the Gender template in Section 4.6.3.6, 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.

4.7 Augmentation

We present an approach for augmenting ContraPro to improve CR. Augmentation systematically expands the data to improve a model's robustness (Kafle et al., 2017). While challenging for NLP, we focus on a narrow problem which lends itself to easier data manipulation. Figure 4.2 shows that our model is capable of modeling the gender of nouns. However, there is a strong prior for translating *it* to *es* and hence little intelligent CR capability. Our goal with the augmentation is to alter the prior and test if this can improve CR in the model.

We augment 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. Table 4.3 provides an additional example. Antecedents and coreferential pronouns are identified using a CR tool (Clark and Manning, 2016a,b). 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 to confidently say that any potential improvements come from the augmentations.

4.7.1 Augmentation Improves Coreference Accuracy

Results for adversarial attacks on ContraPro and ContraCAT are independent.

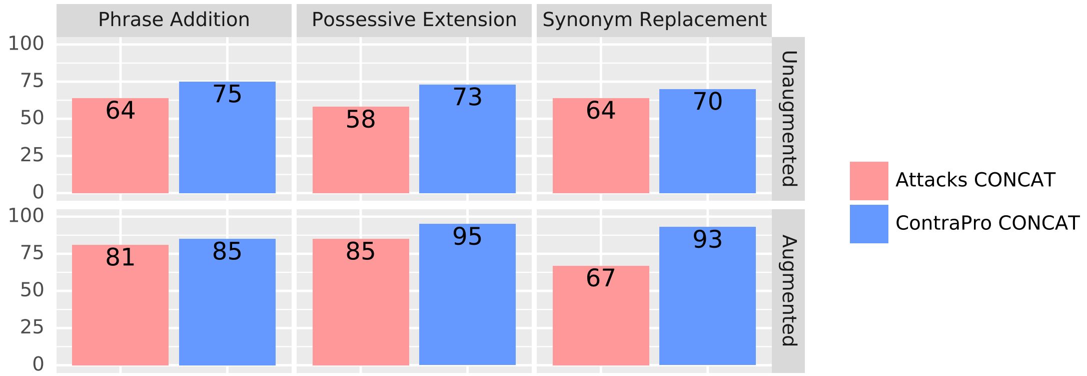

Figure 4.3: Results comparing unaugmented and augmented CONCAT on ContraPro and same 3 attacks as in Figure 4.1. Results with non-augmented CONCAT are the same as Figure 4.1.

4.7.1.1 Adversarial Attacks

AFA provides large improvements, scoring 85.3% on ContraPro (Figure 4.3). Since the datasets themselves are slightly different due to the augmentation, we must recompute the baseline. The AFA baseline (fine-tuning on the augmentation candidates only) is higher by 1.94%, presumably because many candidates consist of coreference chains of “it” and the model learns they are important for coreferential pronouns. This improvement in the baseline is small compared to AFA improvements in the full models.

Prediction accuracy on *er* and *sie* is substantially increased, suggesting that the augmentation removes the strong bias towards *es*. Although, the adversarial attacks lower AFA scores, in contrast to CONCAT, the model is more robust and the accuracy degradation is substantially lower (except on the synonym attack). We experiment with different learning rates during fine-tuning and present results with the LR that obtain the best baseline ContraPro score. Furthermore, CONCAT and AFA obtain 31.5 and 32.2 BLEU on ContraPro, showing that this fine-tuning procedure, which is tailored to pronoun translation, does not lead to any degradation in overall translation quality.

4.7.1.2 Templates

The prior over gender pronouns less concentrated on *es*. This provides for a more even distribution on the **Position** and **Role Prior** template.

The augmented model has higher accuracy on **Markable Detection**, improving by 27.6%. Results for the templates are in Figure 4.4.

No improvements are observed on the World Knowledge template. Pleonastic cases are still accurate, although not perfect as with CONCAT. The Event template identifies a systematic issue with our augmentation. We presume this is due to the CR tool marking cases where *it* refers to events. We do not apply any filtering and augment these cases as well, thus creating wrong examples (an event reference *it* cannot be translated to *er* or *sie*). As a result, the scores are lower compared to

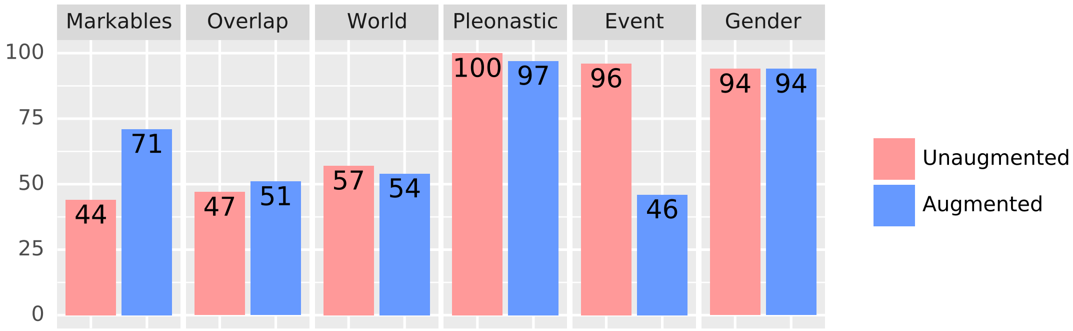

Figure 4.4: ContraCAT results with unaugmented and augmented CONCAT. We speculate that readjusting the prior over genders in augmented CONCAT explains the improvements on Markables and Overlap.

CONCAT. This issue with our model is not visible on ContraPro and the adversarial attacks results. In contrast, the Event template easily identifies this problem.

AFA has a similar accuracy to 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.

4.8 Our Dataset in Context

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, 2018; Miculicich et al., 2018; Voita et al., 2019; Maruf et al., 2019).

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 accuracy scores, precipitating our new dataset.

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 a 1:1 correspondence between pronouns in different languages and that 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.

4.9 Implications for Machine Translation and Automation

In this work, we study how and to what extent CR is handled in context-aware NMT. This work shows 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, 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 to see if simple data modifications can improve model accuracy. This methodology illustrates the dependence on data by models, and strengthen our claims that low-cost data **generation** techniques are creating datasets that approximate rather than solve NLP tasks. Having identified limitations in existing models, we argue for concrete data extensions for coreference resolution. This methodology—creating an adverserial dataset which tests the understanding of a model—can be applied to most NLP tasks.

This project introduces using an **expert**, in this case a native German speaker, in designing the dataset. However, we use templates rather than experts to **automatically** scale the size of the dataset. While we can create *large* datasets, they end up (literally) formulaic. *Solving* tasks like coreference, rather than just noting shortcomings of current datasets, will require building complex and nuanced datasets that allow a model to earn the edge cases of the task. These datasets will ultimately have to built by humans and not **automation**: can the **crowd** be a reliable source of language?

Template Target	**Example**
Priors	
Grammatical Role	The ***cat*** ate the ***egg***. It (*cat*/*egg*) was big.
Order	I stood in front of the ***cat*** and the ***dog***. It (*cat*/*dog*) was big.
Verb	Wow! She unlocked it.
Markable Detection	
Filter Humans	The ***cat*** and the *actress* were happy. However it (*cat*) was happier.
Coreference Resolution	
Lexical Overlap	The ***cat*** ate the apple and the *owl* drank the water. It (*cat*/ *dogFir*) ate the apple quickly.
World Knowledge	The ***cat*** ate the *cookie*. It (*cat*) 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(*cat*) was big. → Ich habe eine Katze gesehen. Sie (*cat*) war groß.

Table 4.2: Template examples targeting different CR steps and substeps. For German, we create three versions with *er*, *sie*, or *es* as different translations of *it*.

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

Bibliography

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.

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 *Empirical Methods on Natural Language Processing*.

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. 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. 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. 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.

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. 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).

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.

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. Association for Computational Linguistics.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202.

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.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

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.

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. Association for Computational Linguistics.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the Association for Computational Linguistics*.

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. 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 multipass 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. 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 Association for Computational Linguistics*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

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.

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.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*.

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

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.

Xiaosen Wang, Hao Jin, and Kun He. 2019. Natural language adversarial attacks and defenses in word level. *arXiv preprint arXiv:1909.06723*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. arXiv preprintado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv: 1609.08144*.

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*.