# Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite

**Prathyusha Jwalapuram**[*], **Shafiq Joty**[*§], **Irina Temnikova**[†], **and Preslav Nakov**[‡]

[*]Nanyang Technological University, Singapore
[§]Salesforce Research Asia, Singapore
[†]Sofia University, Bulgaria
[‡]Qatar Computing Research Institute, Doha, Qatar
[*]{jwal0001,srjoty}@ntu.edu.sg
[†]irina.temnikova@gmail.com
[‡]pnakov@hbku.edu.qa

## Abstract

The neural revolution in machine translation has made it easier to model larger contexts beyond the sentence-level, which can potentially help resolve some discourse-level ambiguities such as pronominal anaphora, thus enabling better translations. Unfortunately, even when the resulting improvements are seen as substantial by humans, they remain virtually unnoticed by traditional automatic evaluation measures such as BLEU, as only a few words end up being affected. Thus, specialized evaluation measures are needed. With this aim in mind, we contribute an extensive, targeted dataset that can be used as a test suite for pronoun translation, covering multiple source languages and different pronoun errors drawn from real system translations, for English. We further propose an evaluation measure to differentiate good and bad pronoun translations. Finally, we conduct a user study and we report correlation with human judgments.

## 1 Introduction

Traditionally, machine translation (MT) has been performed at the level of individual sentences, i.e., in isolation from the rest of the document. This was due to the nature of the underlying frameworks: word-based (Brown et al., 1993), then phrase-based (Koehn et al., 2003), syntactic (Galley et al., 2004), and hierarchical (Chiang, 2005). While there have been attempts to model the context beyond the sentence level, e.g., looking at neighboring sentences (Carpuat and Wu, 2007; Chan et al., 2007) or even at the entire document (Hardmeier et al., 2012), these approaches were still limited by the underlying framework.

Then, along came the neural revolution. Thanks to the attention mechanism, neural translation models such as sequence-to-sequence (Bahdanau et al., 2015) and the Transformer (Vaswani et al., 2017) could model much broader context.

While initially translation was still done in a sentence-by-sentence fashion, researchers soon realized that going beyond that has become easier, and recent work has successfully exploited this (Bawden et al., 2018; Voita et al., 2018). This is an exciting research direction as it can help address inter-sentential phenomena such as anaphora, gender agreement, lexical consistency, and text coherence, to mention just a few.

Unfortunately, going beyond the sentence level typically yields very few changes in the translation output, and even when these changes are seen as substantial by humans, they remain virtually unnoticed by typical MT evaluation measures such as BLEU (Papineni et al., 2002), which are known to be notoriously problematic for the evaluation of discourse-level aspects in MT (Hardmeier, 2014).

The limitations of BLEU are well-known and have been discussed in detail in a recent study (Reiter, 2018). It has long been argued that as the quality of machine translation improves, there will be a singularity moment when existing evaluation measures would be unable to tell whether a given output was produced by a human or by a machine. Indeed, there have been recent claims that human parity has already been achieved (Hassan et al., 2018), but it has also been shown that it is easy to tell apart a human translation from a machine output when going beyond the sentence level (Läubli et al., 2018). Overall, it is clear that there is a need for machine translation evaluation measures that look beyond the sentence level, and thus can better appreciate the improvements that a discourse-aware MT system could potentially bring.

Alternatively, one could use diagnostic test sets that are designed to evaluate how an MT system handles specific discourse phenomena (Bawden et al., 2018; Rios et al., 2018). There have also been proposals to use semi-automatic measures and test suites (Guillou and Hardmeier, 2018).

Here we propose a targeted dataset for machine translation evaluation with a focus on anaphora. We further present a specialized evaluation measure trained on this dataset. The measure performs pairwise evaluations: it learns to distinguish good vs. bad translations of pronouns, without being given specific signals of the errors. It has been argued that pairwise evaluation is useful and sufficient for machine translation evaluation (Guzmán et al., 2015, 2017). In particular, Duh (2008) has shown that ranking-based evaluation measures can achieve higher correlations with human judgments, as ranking judgments are easier to obtain from human judges and are also easy to use in training, while also directly achieving the purpose of comparing two systems.

Note that while it may be possible to rank translations using strong pre-trained conditional language models such as GPT (Radford et al., 2018), all kinds of errors would influence the score, and it would not be targeted towards a specific source of error, such as anaphora here. Our model provides a way to do this, and we demonstrate that it indeed focuses on the translation of pronouns.

Although our pronoun test suite naturally consists of the source text paired with a reference human translation, our pronoun evaluation measure is generally independent of the source language. Moreover, we use real machine translation output, which may contain various types of errors. Our contributions are as follows:

- We create a dataset for pronoun translation that covers multiple source languages and various target English pronouns.

- We propose a novel evaluation measure that differentiates good pronoun translations from bad ones irrespective of the source language they were translated from.

- Unlike previous work, both the dataset and the model are based on actual system outputs.

- Our evaluation measure achieves high agreement with human judgments.

We make both the dataset and the evaluation measure publicly available at https://ntunlpsg.github.io/project/discomt/eval-anaphora/.

## 2 Related Work

Previous work on discourse-aware machine translation and MT evaluation has targeted a number of phenomena such as anaphora, gender agreement, lexical consistency, and coherence. In this work, we focus on pronoun translation.

Pronoun translation has been the target of a shared task at the DiscoMT and WMT workshops in 2015-2017 (Hardmeier et al., 2015; Guillou et al., 2016; Loáiciga et al., 2017). However, the focus was on cross-lingual pronoun prediction, which required choosing the correct pronouns in the context of an existing translation, i.e., this was not a realistic translation task. The 2015 edition of the task also featured a pronoun-focused translation task, which was like a normal MT task except that the evaluation focused on the pronouns only, and was performed manually. In contrast, we have a real MT evaluation setup, and we develop and use a fully automatic evaluation measure.

More recently, there has been a move towards using specialized test suites specifically designed to assess system quality for some fine-grained problematic categories, including pronoun translation. For example, the PROTEST test suite (Guillou and Hardmeier, 2016) comprises 250 pronoun tokens, used in a semi-automatic evaluation: the pronouns in the MT output and in the reference are compared automatically, but in case of no matches, manual evaluation was required. Moreover, no final aggregate score over all pronouns was produced. In contrast, we have a much larger test suite with a fully automatic measure.

Another semi-automatic system is described in (Guillou et al., 2018). It focused on just two pronouns, *it* and *they*, and was applied to a single language pair. In contrast, we have a fully automated evaluation measure, we handle many English pronouns, and we cover multiple source languages.

Bawden et al. (2018) presented hand-crafted discourse test sets to test a model's ability to exploit previous source and target sentences, based on 200 contrastive pairs, one with a correct and one with a wrong pronoun translation. This alleviates the need for an automatic evaluation measure as one can just count how many times the MT system has generated the correct pronoun. In contrast, we use texts from pre-existing MT evaluation datasets, we do not require them to be in contrastive pairs, and we have a fully automated evaluation measure; we also use larger datasets.

Müller et al. (2018) also used contrastive translation pairs, mined from a parallel corpus using automatic coreference-based mining of context, thus minimizing the risk of producing wrong contrastive examples that are both valid translations. Yet, they did not propose an evaluation measure.

Finally, there have been pronoun-focused automatic machine translation evaluation measures. Two important examples include APT (Miculicich Werlen and Popescu-Belis, 2017) and Auto-PRF (Hardmeier and Federico, 2010). Both measures require alignments between the source, the reference and the system output texts for evaluating the pronoun translations. However, automatic alignments are noisy; Guillou and Hardmeier (2018) have shown that improvements using heuristics are not statistically significant. They also found low agreement between these measures and human judgments, primarily due to the possibility of many translation choices per pronoun. APT also uses a predetermined list of 'equivalent pronouns', obtained for specific pronouns based on a French grammar book and verified through probability counts. This list is used to weigh pronouns that are not exact matches, and the accuracy of the pronoun translations is calculated accordingly. Miculicich Werlen and Popescu-Belis (2017) collect such a list for English-French for the pronouns *it* and *they*. This limits the evaluation measure both in terms of the language and also of the pronouns it is applicable to. In contrast, our framework requires only two candidate translations of the same text as input for comparison: this could be a reference vs. a system translation, or a comparison between two candidate translations (see Section 5.5).

## 3 Dataset Generation

We automatically generated our dataset, which we used to build a pronoun test suite and to train a pronoun evaluation model. In order to avoid generating synthetic data that may not necessarily represent a difficult context (for an MT system to correctly translate the pronouns), we used data from actual system outputs submitted for the WMT translation tasks in 2011–2015 and 2017 (Callison-Burch et al., 2011, 2012; Bojar et al., 2013, 2014, 2015, 2017). Using such data means that what is essentially a conditional language model solution, such as the one used by Bawden et al. (2018), has already failed on these examples.

**Original French input**: *Il était créatif, généreux, drôle, affectueux et talentueux, et il va beaucoup **me** manquer.*

**Reference translation**: ***He** was creative, generous, funny, loving and talented, and **I** will miss him dearly.*

**MT system translation**: ***It** was creative, generous, funny, affectionate and talented, and **we** will greatly miss.*

**Generated noisy example 1**: ***It** was creative, generous, funny, loving and talented, and I will miss him dearly.*

**Generated noisy example 2**: ***He** was creative, generous, funny, loving and talented, and **we** will miss him dearly.*

Figure 1: Noisy examples generated by substituting an MT-generated pronoun in the reference translation.

In particular, we aligned the system outputs with the reference translation using an automatic alignment tool (Dyer et al., 2013), and we found examples in which the pronouns did not match the reference translation. This process yielded potentially noisy data, as the alignments are automatic and thus not always perfect.

### 3.1 User Study

In order to ensure that the mismatched pronouns are not equally good translations in the given context, we conducted a user study on a subset of the data. To focus the study on pronouns and to remove the influence of other MT errors, we generated a noisy candidate by replacing the correct pronoun in the reference with the aligned (potentially) incorrect pronoun from the system output. We did this for each differing pronoun, so that the difference between the reference and the noisy version is one pronoun only (see Figure 1).

Our goal was to find pronoun pairs, e.g., *he-it* in Figure 1, where there is high agreement that the reference is the correct translation, so that we can automatically classify it as a positive example and the MT output as a negative one. The study participants were fluent in English and were native speakers of Chinese, Russian, French, or German. They were shown the source and two candidate translations (the reference and its noisy version) in random order. The relevant sentence was shown in bold, with the pronoun highlighted. Two previous sentences were given as a context; see Figure 2.

Figure 2: Our annotation framework.

| Language Pair | Nb. of Ann. | Avg% Ref | AC1 (Incl. Ties) | AC1 (Excl. Ties) |
|---|---|---|---|---|
| Russian→English | 3 | 80.2 | 0.82 | 0.92 |
| French→English | 2 | 83.9 | 0.86 | 0.96 |
| German→English | 2 | 84.3 | 0.89 | 0.97 |
| Chinese→English | 3 | 86.0 | 0.86 | 0.91 |
| - - Only English | 3 | 85.3 | 0.84 | 0.92 |

Table 1: Inter-annotator agreement.

We asked the study participants to choose the text with the better pronoun. They were allowed to choose (*a*) candidate A, (*b*) candidate B, (*c*) equivalent translations (tie), (*d*) "neither is correct", and (*e*) "invalid candidates", i.e., the highlighted words are not pronouns or are the wrong pronouns due to misalignment. We excluded from further consideration all examples marked as (*d*) or (*e*). Each participant annotated a total of **500 examples** per language pair. Statistics[1] about the annotation process are given in Table 1. We also report the proportion of cases where the participants preferred the reference translation over the noisy version (see **Avg%Ref**).[2] We can see that there is high agreement for all language pairs, ranging from 0.82 to 0.89. The ties seem to be the major source of disagreement: excluding them yields agreements in the range of 0.91–0.97.

In order to measure the effect, if any, that the source text might have on the annotator's choices, we also conducted a study without providing them the source text. We did this for the texts from the Chinese→English study. The participants were only shown the English texts: the reference vs. the noisy sentence, with the context as before. The results for this study are shown on the last line of Table 1. We can see that the agreement for this English-only setup is also fairly high (0.84); the overall agreement between all six participants (three from Chinese→English and three from only English) is 0.85. We observe very similar agreement of 0.91 (Chinese→English) and 0.92 (Only English) when the ties were excluded, with the overall agreement being 0.90 between the six participants. However, further analysis showed that although the two groups disagreed on about 10% of the examples, only 2% of the examples were common to both groups, showing that the sources of disagreement between the two groups are different. It can be that having the source context helps disambiguate the other 8% of the cases, while also introducing ambiguity that does not seem to be an issue for the participants who saw the English texts only. See Figure 1 for an example where the source is helpful; here, noisy example 2 would be acceptable, except that the original French text uses a singular pronoun. However, the disagreements form a small part of the dataset; we also filtered out all pronoun pairs with low agreements from further use. Figure 3 shows one such low-agreement example.

---

[1] Due to the nature of the dataset, the human annotators are always more likely to choose the reference as the better candidate, which yields a skewed distribution of the annotations; traditional correlation measures such as Cohen's kappa are not robust to this, and thus we report the more appropriate Gwet's AC1/gamma coefficient.(Gwet, 2008).

[2] High agreement could also mean that the participants consistently picked the noisy version as the better choice.

**Source Text**: 不过现在，她只想享受当下。我不想说这是我的最后一场比赛。这会给我带来太大的压力。

**Reference translation**: *For now she just wants to enjoy the moment. I didn't want to say **this** was my last race. That would have meant too much pressure.*

**Noisy candidate**: *For now she just wants to enjoy the moment. I didn't want to say **that** was my last race. That would have meant too much pressure.*

Figure 3: Low-agreement example: Chinese-English.

## 3.2 Pronoun Test Suite for MT Systems

The source sentences can also be used as a **test suite** for MT systems to check their pronoun translations: it can be considered to be a challenging, diagnostic test set for pronoun translation, covering a range of errors such as gender (he/she→it), number (they→it), animacy (who→which), syntactic role (e.g., subject/object: he→him), and others; see the Appendix for a complete list.

The WMT test sets come from news articles; the context is available, and thus the test suite is particularly suitable for discourse-level MT systems. It is available for each source language for which English translations are generated in the WMT tasks, e.g., Czech, French, German, etc. (Table 2).

The corresponding noisy versions of the reference are also generated, but are somewhat noisy. However, the subset used in our study is curated in some sense as human judgments are available. This data can serve as a more refined test suite: not only useful for checking the agreement with human judgments, but also for identifying equivalent pronoun translations in context, as the data is also annotated for ties.

| Source Language | Test Data from WMT Years | Unique Source Contexts |
|---|---|---|
| German | 2011-2015,17 | 7,823 |
| Czech | 2011-2015,2017 | 6,713 |
| French | 2011-2015 | 4,659 |
| Russian | 2013,2014,2017 | 4,513 |
| Spanish | 2011-2013 | 4,417 |
| Finnish | 2015,2017 | 1,551 |
| Turkish | 2017 | 1,372 |
| Hindi | 2014 | 921 |
| Chinese | 2017 | 696 |
| Latvian | 2017 | 652 |

Table 2: Pronoun test suite for MT systems: English as a target and various languages as a source.

## 4 The Evaluation Measure

While *diagnostic datasets* allow us to evaluate MT systems with respect to specific discourse-level phenomena, an automatic discourse-aware *evaluation measure* is useful not only for evaluation but also for tuning MT systems. Moreover, an evaluation measure that only looks at the target language (which is computationally feasible, even if not ideal, as our study above has shown) offers additional benefits; we can train it for a specific target language without requiring a separate dataset for each source-target language pair. Below, we propose such a measure for pronoun translation.

Let $R = (C_r; r)$ and $S = (C_s; s)$ denote a *reference* and a *system* tuple pair containing a reference and a system translation, $r$ and $s$, along with a context of previous sentences, $C_r$ and $C_s$, respectively. Note that $C_r$ and $C_s$ can contain the same sentences or different sentences, or be empty in case no context is provided. Given a training set $\mathcal{D} = \{(R_i, S_i)\}_{i=1}^{N}$ containing $N$ such tuple pairs, our aim is to learn an evaluation measure that can rank any unseen translation pair $(R, S)$ with respect to the correct use of pronouns.[3] In Section 3, we described how such datasets can be collected opportunistically without recourse to expensive manual annotation.

Figure 4 shows our proposed framework to evaluate MT outputs with respect to pronouns. The inputs to the model are sentences (with or without context $C_r$ and $C_s$): $R$ and $S$. Each input sentence is first mapped into a set of word embedding vectors of dimensionality $d$ by performing a lookup in the shared embedding matrix $E \in \mathbb{R}^{v \times d}$ with vocabulary size $v$. $E$ can be initialized randomly or with any pre-trained embeddings such as GloVe (Pennington et al., 2014), or contextualized word vectors such as ELMo (Peters et al., 2018a).

In case of initialization with GloVe vectors, we use a BiLSTM (Hochreiter and Schmidhuber, 1997) layer to get a representation of the words that is encoded with contextual information. Let $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ denote an input sequence, where $\mathbf{x}_t$ is the $t^{th}$ word embedding vector of the sequence. The LSTM recurrent layer computes a compositional representation $\mathbf{k}_t$ at every time step $t$ by performing nonlinear transformations of the current input $\mathbf{x}_t$ and the output of the previous time step $\mathbf{k}_{t-1}$.

---

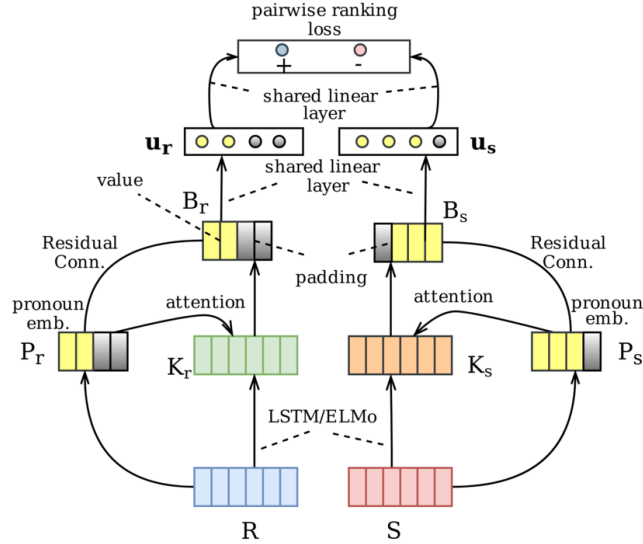[3]Here $R$ (or $S$) can be a reference or a system translation.

Figure 4: Our proposed framework to differentiate good pronoun translations from bad ones in context.

In a BiLSTM, we get the representation $\overrightarrow{\mathbf{k}_t}$ by processing the sequence in the forward direction, and the representation $\overleftarrow{\mathbf{k}_t}$ by processing the sequence in the backward direction. The final representation $\mathbf{k}_t$ of a word is the concatenation of these two representations, i.e., $\mathbf{k}_t = [\overrightarrow{\mathbf{k}_t}; \overleftarrow{\mathbf{k}_t}]$.

With ELMo initialization, the word vectors obtained are used directly. ELMo uses stacked biLSTM encoder and gives very powerful contextualized word representations learned from large corpora by optimizing a bi-directional language modeling loss. The ELMo representations already capture morphological, syntactic and contextual semantic features (Peters et al., 2018b).[4]

Let $K_r$ and $K_s$ be the matrices whose rows represent the word representations of $R$ and $S$, respectively (obtained either from Bi-LSTM or directly from ELMo). From these representations, we extract the representations of the pronouns in the target sentence (from $r$ and $s$; not from the contexts). Let $P_r$ and $P_s$ be the matrices whose rows represent the contextualized word representations of the pronouns in $r$ and $s$, respectively. We use zero-padding (shown as shaded boxes) to make $P_r$ and $P_s$ fixed-length. We then use *scaled multiplicative attention* (Vaswani et al., 2017) to compute a contextual representation for the pronouns in $r$ and $s$. Specifically, we consider the rows of $P_r$ (*resp.* $P_s$) as *query* vectors, the rows of $K_r$ (*resp.* $K_s$) as *key* and *value* vectors, and the matrix $P_r$ (*resp.* $P_s$) to attend over $K_r$ (*resp.* $K_s$).

We use residual connection and layer normalization to get pronoun representations $B_r$ and $B_s$:

$$B'_r = \mathcal{S}(\frac{P_r K_r^T}{\sqrt[2]{d}})K_r; \quad B'_s = \mathcal{S}(\frac{P_s K_s^T}{\sqrt[2]{d}})K_s \quad (1)$$

$$B_r = \text{LayerNor}(P_r + B'_r); \quad B_s = \text{LayerNor}(P_s + B'_s) \quad (2)$$

Note that $B_r$ and $B_s$ contain a $d$-dimensional vector for each query (pronoun) vector (and *zero* vectors due to padding). We pass these vectors through a **shared** linear layer parameterized by $\mathbf{z} \in \mathbb{R}^d$ to obtain a score for each pronoun. This yields vectors $\mathbf{u}_r$ and $\mathbf{u}_s$ for the reference and for the system translations:

$$\mathbf{u}_r = B_r \mathbf{z}; \quad \mathbf{u}_s = B_s \mathbf{z}; \quad (3)$$

A final **shared** linear layer parameterized by $\mathbf{w}$ converts these vectors to contrastive scores, yielding a (positive) score $y_r$ for the reference and a (negative) score $y_s$ for the system translation:

$$y_r = \mathbf{u}_r^T \mathbf{w}; \quad y_s = \mathbf{u}_s^T \mathbf{w}; \quad (4)$$

We then use the scores in a pairwise ranking loss (Collobert et al., 2011) to find model parameters that assign a higher score to $y_r$ than to $y_s$. We minimize the following ranking objective:

$$\mathcal{L}(\theta) = \max\{0, 0.1 - y_r + y_s\} \quad (5)$$

---

[4] We also tried an ELMo-initialized BiLSTM, but it did not perform well while increasing model complexity.

Note that the network shares all of its parameters ($\theta$) to obtain $y_r$ and $y_s$ from a pair of inputs $R_i = (C_r, r)$ and $S_i = (C_s, s)$. Once trained, it can be used to score any input independently.

# 5 Experiments

Below, we describe our data, the experimental setup, and the evaluation results.

## 5.1 Data

We first created a set of commonly confused pronoun pairs. Using the data from the study, we calculated the inter-annotator agreement for each pair of a reference/correct pronoun and a system translation/incorrect pronoun. We excluded the pairs with low agreement ($<0.8$) or for which the system output was chosen as the correct translation more often. Pairs with low agreement are essentially cases where the annotators cannot agree that the reference translation is better. The source of ambiguity in this case is that the system translation is not absolutely wrong (see Figure 3); therefore, these cases may not be so critical to correct. The remaining pairs are, with a fairly high confidence, positive–negative (correct–incorrect) pairs.

Next, we filtered the WMT data, only keeping sentences with these pronoun pairs. This yielded 97,461 reference translation (positive text) — unique system output (negative text) pairs for training, taken from WMT11,12,13,15 (Table 3).

The development data collected from WMT14 system outputs has 5,727 unique system translations and 6,635 unique noisy candidate pairs.[5]

For testing, we used the annotated data from the user study, generated from a subset of WMT17 system translations (except for French, which is from the discussion forum test set from WMT15, not overlapping with the training data). There are 500 unique noisy-reference pairs per source language, a total of 2,000.

| Data | Source | #Unique Pairs |
|------|--------|---------------|
| Training | WMT11-13,15 | 97,461 |
| Development | WMT14 | 5,727 |
| Development (noisy) | WMT14 | 6,635 |
| Test (noisy) | WMT15,17 | 2,000 |

Table 3: Statistics about our dataset.

---

[5] The number of unique noisy candidates exceeds that of unique system translations because a separate noisy candidate was generated for each error in a system translation.

## 5.2 Experimental Setup

We evaluated the models in terms of accuracy, i.e., the proportion of times the model scored the reference translation higher than the system/noisy output, and we report results using either GloVe (Pennington et al., 2014) or ELMo (Peters et al., 2018a). We conducted a number of experiments, training and testing under different conditions:

**No Context (NC).** The reference ($R$) (or the noisy reference $R'$) and the system ($S$) translations go through ELMo or BiLSTM, without contextual information $C_r/C_s$ (i.e., query = $P_r/P_s$, key, value = $R = r/S = s$);

**With Context:** The reference ($R$) (or the noisy reference $R'$) and the system ($S$) translation representations include two previous sentences as a context. The context can be further categorized as

(i) **Respective Context (RC)**: $R$ includes its own reference context $C_r = r_{-2}r_{-1}$, and $S$ includes its own respective system context $C_s = s_{-2}s_{-1}$ (query = $P_r/P_s$, key, value = $R = r_{-2}r_{-1}r/S = s_{-2}s_{-1}s$);

(ii) **Common Reference Context (CRC)**: The context for $R$ and $S$ includes the same reference context $C_r = C_s = r_{-2}r_{-1}$ for both (query = $P_r/P_s$, key, value = $R = r_{-2}r_{-1}r/S = r_{-2}r_{-1}s$);

We perform the evaluation in two ways:

(a) **R vs. S**: Testing over pairs of reference ($R$) and system translation ($S$) texts;

(b) **R vs. R'**: Testing over pairs of reference ($R$) and noisy candidate ($R'$) texts.

**Baseline.** For a baseline performance, we simply take the average of the extracted pronoun representations in $P_r$ and $P_s$, and we convert them to pairwise scores through linear layers. The baseline is also evaluated with and without context.

| Exp | Context Setting | Test | Acc. (Glove) | Acc. (ELMo) |
|-----|-----------------|------|--------------|-------------|
| 1 | NC-Baseline | **R** vs. **R'** | 69.12 | 85.80 |
| 2 | NC | **R** vs. **R'** | 68.97 | 88.04 |
| 3 | NC | **R** vs. **S** | 79.67 | 89.09 |
| 4 | RC-Baseline | **R** vs. **R'** | 69.07 | 85.80 |
| 5 | RC | **R** vs. **R'** | 67.88 | 87.90 |
| 6 | CRC-Baseline | **R** vs. **R'** | 69.16 | 86.66 |
| 7 | CRC | **R** vs. **R'** | 68.93 | 89.11 |
| 8 | CRC | **R** vs. **S** | 77.87 | 90.69 |

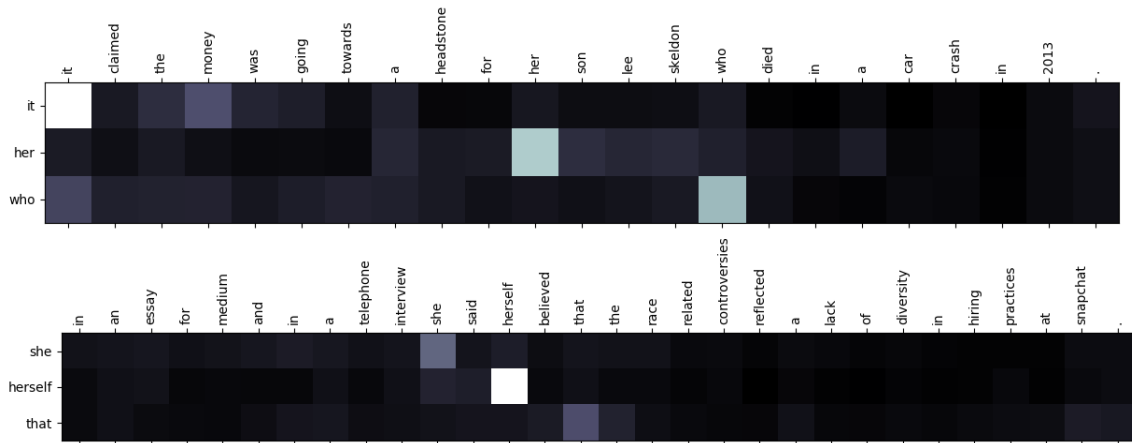Table 4: Results on WMT14 with different contexts.

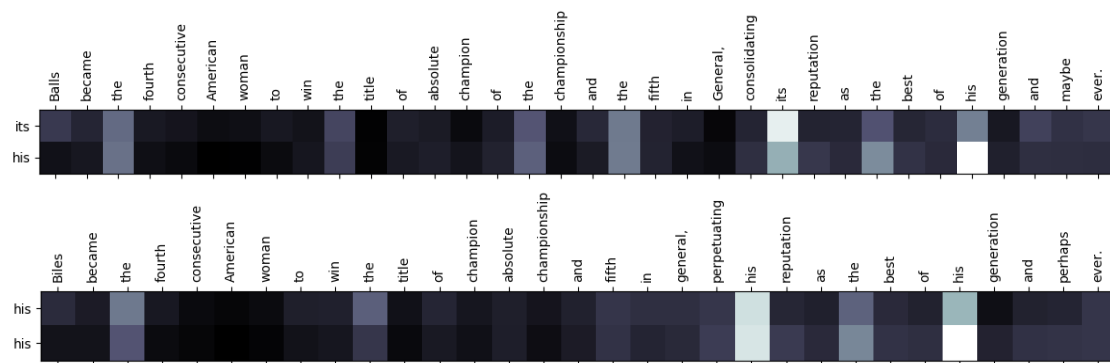Figure 5: Attention maps for some noisy candidates from the test set.



Figure 6: Attention maps for two candidate system translations of the same text (best viewed in color)

## 5.3 Results and Analysis

The first three experiments in Table 4 show results for the 'No Context' setting. These are on the noisy data (Exp. 1-2) and are indicative of how sensitive our model is to pronouns, since $R$ and $R'$ differ in a single pronoun. The results on the reference/system translation pair $R$ and $S$ (Exp. 3) is indicative of the performance in a real use case. In the case of the GloVe-BiLSTM model, using attention instead of averaging pronouns does not help. However, using ELMo greatly improves the accuracy of the baseline and also yields an improvement when using attention. ELMo can model syntactic and semantic information and can thus improve co-reference resolution (Peters et al., 2018a), which could be a contributing factor.

The second part of our experiments (Exp. 4-5) concerns the addition of contextual information. When the respective contexts are added, there is no improvement (even a drop in case of GloVe-BiLSTM); quite possibly, other differences in the text make it harder for the model to focus on the pronoun errors.

To offset this issue, we use a common context for both the reference and the system sentences, taken from the reference (Exp. 6-8). Training the model with a common reference context (CRC) yields marginal difference in the GloVe-BiLSTM model, but the ELMo model improves with the addition of context. Our experiments[6] show that ELMo is quite powerful at capturing contextual information, even as the context size grows.

Finally, we test our *ELMo - common reference context* model on the held-out dataset used for the user study. Table 5 shows the results. The accuracy of the system is lower on the study dataset. Note that since the training data was filtered based on the pronoun pairs with high agreement from the study, the study dataset contains pronoun pairs with low agreement that were not seen during training. We also calculate the agreement with human judgments excluding ties since our model does not handle ties. The overall agreement with the human judgments remains high.

---

[6]We also ran experiments with BERT (Devlin et al., 2019) and found that it performed on par with ELMo.

| Language | Acc.(ELMo) | AC1 Agr. |
|---|---|---|
| Russian→English | 79.4 | 0.80 |
| French→English | 82.0 | 0.84 |
| German→English | 81.6 | 0.83 |
| Chinese→English | 82.4 | 0.83 |
| - - Only English | —- | 0.83 |
| **Overall (average)** | 81.35 | —- |

Table 5: **Acc**uracy and AC1 **Agr**eement for the ELMo-based model predictions on the study dataset.

## 5.4 Pronoun-wise Analysis

We performed a pronoun-wise analysis of the results in Table 5. The model scored the noisy version higher than the reference in about 19% of the cases. Of these, 46% were pronoun pairs that were not seen during training.

Of the remaining pronoun pairs that were seen during training, the main source of errors (over 11%) were cases when *that* in the reference was replaced by *it* in the noisy version (*"We always tell victims not to pay up; that/it simply exacerbates the problem", explains Kleczynski*). The noisy candidate was scored higher about 28% of the time, out of 79 samples. The next highest source of errors was the reference–noisy pair of *it–she* at 10%.

In contrast, the best-performing pair was when a *he* in the reference was replaced with an *it* in the noisy version (*He/It risked everything to save other people's lives.*). The reference was scored higher than the noisy candidate 95% of the time, out of 135 samples. The next highest performer was the reference–noisy pair of *his–its*, which was correctly scored 86% of the time.

This performance follows the distribution of the pronoun pairs seen during training. The *he–it* and *his–its* pairs together account for over 12% of the training data, while the *that–it* and *it–she* pairs together form only 3.7%. Since the distribution of the pronoun pairs in the training data is itself based on the distribution of errors in the system translations, the model performs best over error cases that occur most often in system translations.

## 5.5 Discussion

As systems participating in WMT improve over the years, our test data is closer to state-of-the-art neural MT, while most of the training data is from statistical systems. This setting allows us to capture a wider range of errors while showing that our model is sensitive to small errors in a fluent output.

Note that since the model was trained on the full system output and also on all pronouns in that output, it has not received any signal about which pronoun is wrong. Yet, we can see from the attention maps in Figure 5 that the model can correctly identify the incorrect pronoun. In Figure 5 (top), the model distinguishes the wrong pronoun *it* (the correct one is *she*), while in Figure 5 (bottom) it correctly finds *herself* as the wrong translation (the correct one is *she*).

We further compare the scores of two system translations for the same sentence from WMT17 Russian-English system outputs (see Figure 6). Here, the correct pronoun to be found is *her*. While one system translates it alternately as *its* and *his* (see Figure 6, top), the other system translates both cases as *his* (see Figure 6, bottom). Our model scores the translation in Figure 6 (bottom) higher than the translation in Figure 6 (top); even though the model highlights both occurrences of *his* as wrong, it ends up believing that having *its* as a translation is worse. We could argue that the translation in Figure 6 (bottom) is better since it maintains the animacy/human aspect, even if the grammatical gender is wrong; moreover, this translation is also consistent.

One could further argue that pronoun-focused automatic machine translation evaluation measures such as APT and AutoPRF are likely to yield the same accuracy/precision-recall for both cases above.

## 6 Conclusions and Future Work

We have presented a new, extensive, targeted dataset for pronoun translation that covers multiple source languages and a wide range of target English pronouns. We have also proposed a novel evaluation measure for differentiating good vs. bad pronoun translations, irrespective of the source language, which achieved high correlation with human judgments.

In future work, we want to handle cases where multiple pronouns are equally suitable in a given context. We would also like to extend the work to other discourse phenomena.

# References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations*, ICLR '15, San Diego, CA, USA.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 1304–1313, New Orleans, LA, USA.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, pages 12–58, Baltimore, MD, USA.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT '13, pages 1–44, Sofia, Bulgaria.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation. In *Proceedings of the Second Conference on Machine Translation*, WMT '17, pages 169–214, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, WMT '15, pages 1–46, Lisbon, Portugal.

Peter E. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 10–51, Montréal, Canada.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 22–64, Edinburgh, Scotland, UK.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 61–72, Prague, Czech Republic.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 33–40, Prague, Czech Republic.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Ann Arbor, MI, USA.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, MN, USA.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 191–194, Columbus, OH, USA.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, GA, USA.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '04, pages 273–280, Boston, MA, USA.

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.

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*, EMNLP '18, pages 4797–4802, Brussels, Belgium.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, WMT '18, pages 570–577, Brussels, Belgium.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 525–542, Berlin, Germany.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 805–814, Beijing, China.

Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez i Villodre, and Preslav Nakov. 2017. Machine translation evaluation with neural networks. *Computer Speech & Language*, 45:180–200.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 2010 International Workshop on Spoken Language Translation*, IWSLT '10, pages 283–289, Paris, France.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings*

of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1179–1190, Jeju Island, Korea.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL-HLT '03, pages 48–54, Edmonton, Canada.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 4791–4796, Brussels, Belgium.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DiscoMT '17, pages 1–16, Copenhagen, Denmark.

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*, WMT '17, pages 17–25, Copenhagen, Denmark.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, WMT '18, pages 61–72, Belgium, Brussels.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

*40th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 311–318, Philadelphia, PA, USA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'18, pages 2227–2237, New Orleans, LA, USA.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 1499–1509, Brussels, Belgium.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, WMT '18, pages 594–602, Belgium, Brussels.

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 30*, NIPS '17, pages 5998–6008.

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*, ACL '18, pages 1264–1274, Melbourne, Australia.