Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information

Niccolò Campolungo
Sapienza University of Rome
campolungo@di.uniroma1.it

Tommaso Pasini
Sapienza University of Rome
pasini@di.uniroma1.it

Denis Emelin
University of Edinburgh, Scotland
d.emelin@sms.ed.ac.uk

Roberto Navigli
Sapienza University of Rome
navigli@diag.uniroma1.it

Abstract

Recent studies have shed some light on a common pitfall of Neural Machine Translation (NMT) models, stemming from their struggle to disambiguate polysemous words without lapsing into their most frequently occurring senses in the training corpus. In this paper, we first provide a novel approach for automatically creating high-precision sense-annotated parallel corpora, and then put forward a specifically tailored fine-tuning strategy for exploiting these sense annotations during training without introducing any additional requirement at inference time. The use of explicit senses proved to be beneficial to reduce the disambiguation bias of a baseline NMT model, while, at the same time, leading our system to attain higher BLEU scores than its vanilla counterpart in 3 language pairs.

1 Introduction

Translating a sentence requires the underlying meaning to be captured and then expressed in the target language. Nonetheless, only little attention has been devoted to studying the actual capabilities of Neural Machine Translation (NMT) approaches of modeling different senses of ambiguous words, with recent work showing that systems tend to be biased towards the most frequent meanings found within the training corpus (Emelin et al., 2020). This phenomenon is hard to measure through classical evaluation metrics, such as the BLEU score (Papineni et al., 2002), as they often rely on word-matching heuristics that fail to capture the disambiguation capabilities of the evaluated systems. Therefore, several efforts have been recently devoted to shed some light and create test beds (Rios Gonzales et al., 2017; Raganato et al., 2019; Emelin et al., 2020; Campolungo et al., 2022) to challenge NMT models. Results show that these models still struggle to deal with highly polysemous words, especially when used to express least frequent senses.

For example, given the sentence "The energy comes from a distant plant.", both Google Translate and DeepL disambiguate[1] *plant* to its sense of *organism* when translating into Italian, and produce the following incorrect sentence "L'energia proviene da una *pianta* lontana.", rather than "L'energia proviene da un *impianto* lontano.", where *impianto* is the translation for the factory meaning of *plant*. This suggests that, even when adequate context is provided (*energy* should be enough to correctly infer the right sense of *plant*), state-of-the-art models might still be biased towards the most frequent meanings found within training data.

Some recent studies have explored how to leverage explicit sense information within NMT models (Rios Gonzales et al., 2017; Pu et al., 2018a; Nguyen et al., 2018). Nevertheless, including such information is not trivial for three main reasons: i) sense-tagged parallel data is scarce; ii) Word Sense Disambiguation (WSD) systems have not been accurate enough until very recently (Blevins and Zettlemoyer, 2020; Barba et al., 2021); and iii) how explicit senses should be incorporated within neural models is not straightforward.

In this paper, we first introduce a novel approach to make up for the paucity of sense annotations in parallel corpora, leveraging a multilingual WSD system to tag parallel sentences and refine its predictions by means of cross-lingual word alignments and information from a multilingual knowledge base. Then, we fine-tune our baseline models on our sense-tagged corpora via a specifically designed loss function, allowing the injection of word-level semantics into the architecture. We evaluate our approach on standard and challenge test sets, showing that it does indeed improve translation accuracy and mitigates the most frequent sense bias.

To summarize, our contributions are manifold:

[1] At the time of writing: January 5th, 2022.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4824 - 4838

1. We put forward a novel approach to produce high-precision sense annotations for parallel data, which we apply to three language pairs.

2. We propose a fine-tuning strategy that lets us inject word-level explicit semantics into Neural Machine Translation models, without introducing any additional requirement at inference time.

3. We show that employing explicit sense tags is beneficial in order both to mitigate the sense bias and to improve the translation quality in terms of BLEU score on standard benchmarks.

4. We present a case study on how a state-of-the-art WSD system compares to an NMT model on disambiguating words within a challenging set for detecting sense bias in MT.

We make all the generated datasets, the code of the model and for the experiments available at https://github.com/sapienzanlp/reducing-wsd-bias-in-nmt.

2 Related Work

Word Sense Disambiguation was first formulated as a computational task by Weaver (1949) in the context of Machine Translation. The two fields then followed parallel paths, with more or less successful attempts over the years to join them back together (Carpuat and Wu, 2005; Vickrey et al., 2005; Carpuat and Wu, 2007). Indeed, while Carpuat and Wu (2005) reported negative results when trying to integrate the prediction of a supervised WSD approach into a Statistical Machine Translation (SMT) model, the same authors, two years later, successfully improved the performance of a phrase-based SMT approach by leveraging a new phrase-based WSD model (Carpuat and Wu, 2007). More recently, Pu et al. (2018a) and Nguyen et al. (2018) proposed systems that successfully leverage sense information in NMT models, although they introduced a heavy requirement, i.e., that of disambiguating the ambiguous words in the sentence prior to generating a translation, which makes them unfeasible in many real-world settings. Lately, contextualized word embeddings have been employed to produce additional back-translated parallel training data via mining sense-specific target sentences, in order to improve handling of infrequent senses (Hangya et al., 2021).

Nevertheless, the proper treatment of lexical ambiguity is still an open problem, with neural models struggling to translate least frequent senses and often relying on spurious correlations among words (Emelin et al., 2020; Raganato et al., 2019; Rios Gonzales et al., 2017). Thus, the disambiguation bias topic has received renewed interest, and several benchmarks have been introduced in the most recent years with the goal of directly measuring the extent to which neural architectures are able to capture word semantics. One of the first of this kind was ContraWSD (Rios Gonzales et al., 2017). In this first attempt to evaluate WSD capabilities of NMT models, the authors built an adversarial test set where source sentences containing an ambiguous word were associated with a correct translation and several incorrect alternatives. These latter were built by replacing the reference translation for the ambiguous word with the translation of one of its other possible meanings. The task measured whether a model ranked the correct translation higher, i.e., it assigned it a higher probability than the adversarial ones. This study provided evaluation data for two language pairs only, i.e., German→English and German→French, and within a few years it became outdated as modern NMT models could easily attain high performances (Emelin et al., 2019). Thus, MuCoW (Raganato et al., 2019) took things a step further and leveraged BabelNet (Navigli and Ponzetto, 2012; Navigli et al., 2021) – a large multilingual knowledge base – and sense embeddings (Camacho-Collados et al., 2016; Mancini et al., 2017) in order to automatically create adversarial translations for five language pairs while also increasing the difficulty of the task itself; however, the fully automatic nature of these challenge sets made them noisy and prone to containing irrelevant challenge samples.

More recently, Emelin et al. (2020) proposed two challenge sets for the English→German pair, one measuring the model sensitivity to most frequent senses and the other estimating, through adversarial injections, its susceptibility to changing a correct sense to a wrong one. In contrast to previous studies, these challenge sets were based on correlations among words in the training set and relied on manually-refined sense clusters, providing an excellent test bed for measuring semantic bias.

Finally, Campolungo et al. (2022) proposed DiBiMT, the first fully manually annotated test set for measuring the disambiguation bias of neural

machine translation models, covering five language combinations, namely, from English to German, Spanish, Italian, Russian and Chinese. In their work, the authors showed that open neural models still exhibit strong semantic biases towards frequent senses, confirming once again the suspicions about this under-explored issue.

Despite all the effort made in putting forward challenging sets of data to test WSD capabilities of NMT models, to the best of our knowledge, only a few approaches (Rios Gonzales et al., 2017; Liu et al., 2018) have been proposed to mitigate this issue, and none of these is effective with modern Transformer-based architectures. Furthermore, while parallel corpora have been exploited to produce sense annotations in the past (Bonansinga and Bond, 2016; Delli Bovi et al., 2017), they were built by utilizing outdated disambiguation approaches that have recently been surpassed by more advanced neural architectures. Indeed, the Word Sense Disambiguation field has received much attention in the last few years, with several supervised approaches (Conia and Navigli, 2021; Blevins and Zettlemoyer, 2020; Barba et al., 2021) and sense embedding models (Loureiro and Jorge, 2019; Scarlini et al., 2020a,b; Wang et al., 2020) performing close to the upper bound limit of the inter-annotator agreement, which finally makes them feasible for inclusion in other downstream tasks, e.g., Machine Translation.

Thus, differently from previous studies in the literature, we focus on closing the gap between these two fields, i.e., Neural Machine Translation and Word Sense Disambiguation, by putting the recent advances in WSD at the service of NMT models. We propose a novel approach, similar to that introduced in Luan et al. (2020), for creating high-quality sense-annotated parallel corpora, and we use this semantic information to regularize an NMT model, making it less biased and capable of producing higher-quality translations.

3 Reducing the Disambiguation Bias in NMT

Neural Machine Translation models are typically trained end-to-end to produce a target translation given a source sentence and, thus, they can only rely on the input context to resolve the ambiguity of polysemous words therein. Being pattern recognition algorithms at heart, these models fall prey to the inherent bias carried by the frequency of co-occurrence of words within parallel sentences, and thus tend to disambiguate words to the sense they most frequently encountered during training, even when the sentence does provide enough context to identify the correct sense. At the same time, Word Sense Disambiguation models, i.e., models specialized in associating a word in context with one of the meanings within a given sense inventory, have recently displayed remarkable results across different benchmarks and languages (Bevilacqua et al., 2021). The time may now therefore be ripe for them to be successfully included into downstream applications such as Neural Machine Translation. However, data that would allow these two worlds to be brought together, i.e., parallel corpora where words are associated with semantic labels, are currently still produced automatically by leveraging outdated approaches to WSD (Delli Bovi et al., 2017).

In what follows, we first provide some preliminary information about resources and tools that we employ in our method (§ 3.1); then, we introduce a new approach for automatically annotating tokens within parallel sentences with sense annotations, i.e., labels explicitly defining their meanings (§ 3.2); finally, we propose a fine-tuning objective for leveraging such annotations in order to mitigate the sense bias while also improving the translation quality overall (§ 3.3). The intuition behind our work is that fixed sense labels describing word senses would help NMT models better encode the underlying meaning of the input sentence, thus generating less biased and overall better translations.

3.1 Preliminaries

We draw sense labels from BabelNet (Navigli and Ponzetto, 2012), a multilingual knowledge base created by merging several semantic resources in different languages such as WordNet (Miller et al., 1990), Wikipedia, Wikidata, etc. BabelNet is structured in synsets, i.e., sets of synonymous senses in different languages. For instance, the synset of $plant^{\text{organism}}$ contains the following lexicalizations: $plant_{EN}$, $pianta_{IT}$, $Pflanze_{DE}$, among others. Additionally, BabelNet provides lemma-to-synsets mappings. For example, the English noun *plant* belongs to the following nominal synsets: *organism*, *industrial plant*, *actor in the audience* and *something placed secretly*.[2] Since BabelNet con-

[2]Synsets bn:00035324n, bn:00046568n, bn:00062800n and bn:00062801n respectively, from https://babelnet.org.

tains millions of synsets, which may make the computation too expensive, we restrict the vocabulary to just those containing at least one English sense from WordNet, as is also done in several other works (Barba et al., 2020; Scarlini et al., 2020b; Bevilacqua and Navigli, 2020).

3.2 Building a Sense-Annotated Parallel Corpus

Let us assume that our running example sentence "The energy comes from a distant plant." appears within a parallel corpus paired with the following Italian translation: "L'energia viene da un impianto lontano.". As we said, by considering the English sentence alone, the word *plant* could take several meanings, among which *organism* and *power plant*. However, among these, only one is shared with its translation *impianto*, i.e., the *power plant* meaning. Therefore, considering the cross-lingual alignment of words may drastically reduce the set of valid meanings, making the disambiguation task much easier. Based on this intuition, given a parallel corpus, we perform the following two steps:

1. **Sense Scoring**, where we employ a WSD system to assign to each content word a distribution over its possible meanings;

2. **Annotation Refinement**, where we compute cross-lingual word alignments to reduce lexical ambiguity and finally assign the most suitable sense to each content word.

Sense Scoring In this step, our goal is to assign to every content word within a sentence a distribution over its possible senses in BabelNet. To this end, given as input a sentence s[3] from a parallel corpus C, we first apply Part-of-Speech tagging and lemmatization to it, then pass it through our WSD system, which returns a distribution over its possible meanings.

Formally, let w_i be a content word in a sentence $s = [w_1, \ldots, w_n]$, and $\sigma(w_i)$ the set of synsets associated with w_i in BabelNet. The WSD system assigns a score $c(S|w_i, s)$ to each synset $S \in \sigma(w_i)$; we denote the synset of w_i with the highest confidence as $S^*_{w_i}$. As a result, each content word in a source or target sentence is associated with a sense distribution. However, applying a WSD system alone may not be sufficient to ensure high-quality annotations, as the application domain may be different from the one of its training set. Therefore, in the next step we take advantage of the translation each sentence is paired with to refine sense annotations.

Annotation Refinement We produce word-level cross-lingual alignments between the source and the target sentences of the parallel corpus: given a pair of parallel sentences (s, t), we compute a list of alignments $\mathcal{A} = \{(w^s_i, w^t_j) | w^s_i \in s, w^t_j \in t\}$. Thus, given an aligned word pair $P = (w^s_i, w^t_j) \in \mathcal{A}$, let $\sigma(P) = \sigma(w^s_i) \cap \sigma(w^t_j)$, i.e., the intersection of synsets that the two words may denote according to BabelNet: we discard annotations for any word pair such that $\sigma(P) = \emptyset \ \vee \ |\sigma(w^s_i)| < 2$. In other words, we retain all the aligned pairs (w^s_i, w^t_j) such that the source word is polysemous and the intersection of their senses is non-empty, thus ensuring higher annotation precision by leveraging the parallelism of words.

Finally, we assign the same synset S^* to both words (w^s_i, w^t_j) in P as follows:

$$
\begin{aligned}
S^* &= S^*_{w^s_i} = S^*_{w^t_j} \\
&= \underset{S \in \sigma(P)}{\operatorname{argmax}} \left(\frac{c(S|w^s_i, s)}{Z_s} + \frac{c(S|w^t_j, t)}{Z_t} \right) \\
Z_s &= \sum_{S \in \sigma(P)} c(S|w^s_i, s) \\
Z_t &= \sum_{S \in \sigma(P)} c(S|w^t_j, t)
\end{aligned}
$$

that is, the synset with the highest combined confidence score after normalizing over $\sigma(P)$, where Z_s and Z_t represent the normalization factors of the probability distributions associated with the synsets of w^s_i and w^t_j, respectively.

3.3 Semantic Injection

Now that we can generate high-quality sense annotations, we describe our fine-tuning method to inject word-level semantics into a Neural Machine Translation model. Ideally, we want the model to benefit from such annotations during training, while not being dependent on them at inference time. To satisfy both these desiderata, we adapt the model's vocabulary to handle synsets as well as subwords, and propose a specific loss that exploits the injected senses to improve the base model's handling of ambiguous words.

[3] s can be either a source or a target sentence.

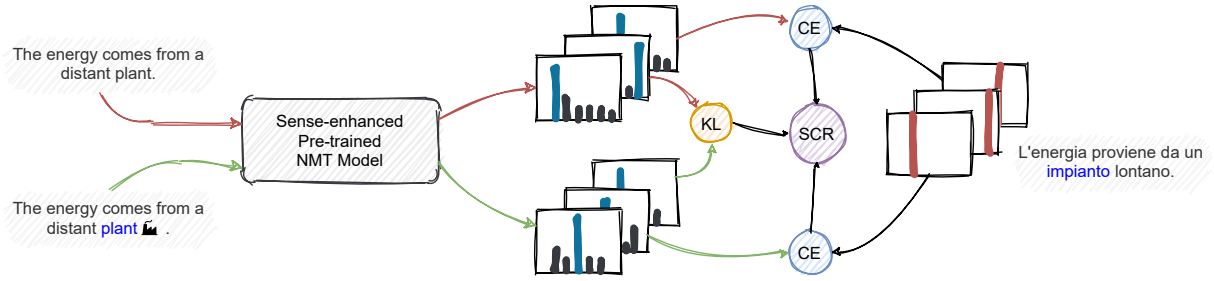

Figure 1: Semantic Consistency Regularization (SCR) fine-tuning. KL stands for Kullback-Leibler divergence; CE stands for Cross Entropy.

Semantically Enhancing Sentences In order to work with concepts, we need a way to represent them. Let us consider once more the sentence "The energy comes from a distant plant.": we rewrite it in order to also include the exact meaning for *plant*, which we computed as described in § 3.2: "The energy comes from a distant plant $plant^{\text{factory}}$".

Formally, given a source sentence s and a word w_i annotated with sense $S^*_{w_i}$, we simply represent w_i as its standard segmentation followed by $S^*_{w_i}$, represented by its sense embedding[4] passed through a linear projection layer (as shown in Figure 2). Additionally, to enforce the connection between the tagged word and its sense annotation, we set the position ids for the word and the sense embedding to the same value, as if they were a single token. This encoding scheme gracefully extends to the whole sentence, yielding the sense-enhanced input representation for a given sentence s.

Semantic Consistency Regularization We hereby propose the Semantic Consistency Regularization (SCR) objective, inspired by MVR (Wang et al., 2021).

Formally, let x' and x'' be two encodings (plain and sense-enhanced) of the same input sentence x and let y be the target sentence, we define SCR as:

$$SCR(\theta) = -\log \mathcal{P}_\theta(y|x') - \log \mathcal{P}_\theta(y|x'') + \mathcal{D}_{\text{KL}}(\mathcal{P}_\theta(y|x') \,||\, \mathcal{P}_\theta(y|x''))$$

where θ is the set of trainable weights, $\mathcal{D}_{\text{KL}}$ is the unidirectional Kullback-Leibler divergence (Kullback and Leibler, 1951) and $\mathcal{P}_\theta(y|x)$ represents an output distribution (a visual representation of SCR is reported in Figure 1).

With this formulation, SCR jointly uses the same sentence with and without sense annotations as two separate inputs: while we train the model to be able to translate both plain and sense-enhanced sentences, by minimizing the divergence between the output distributions we also force the model to transfer the sense information from the sense-enhanced input to the plain input, much like in a self-distillation process. At the same time, we still maintain the model's capability of translating without sense annotations, thus dropping their requirement at inference time.

[4]Details are provided in § 4.

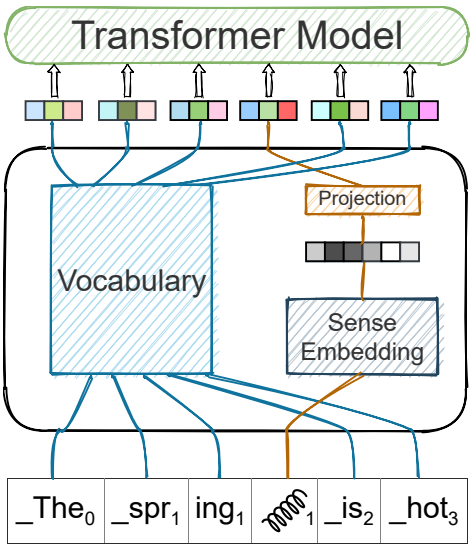

Figure 2: Sense injection mechanism. Subscripts represent the position id associated with the subword.

4 Experimental Setup

4.1 Our Model

We employ as underlying model the standard Transformer architecture (Vaswani et al., 2017), with 6 encoder and 6 decoder layers.[5] Note that, while SCR can be applied to any pre-trained model, we retrain one from scratch because most of the other models available online use part of our test data as

[5]We use randomly-initialized MarianMT models available in HuggingFace's transformers library (Wolf et al., 2020) for easier comparability with their trained versions.

their training data (see § 4.2). Additional details about training configuration and hyperparameters are provided in § A.3.

Fine-tuning with SCR Additionally, to jump-start the model's capabilities, we encode synsets not as randomly initialized learnable vectors (e.g., by extending the vocabulary), but with frozen pre-trained sense embeddings projected into the model's input space by means of a linear layer, the only additional learnable component of the model (*Projection* in Figure 2), which is dropped after the fine-tuning stage. As pre-trained sense embeddings we use ARES (Scarlini et al., 2020b), since they provide multilingual representations for each synset in our vocabulary. We study the impact of this choice in § 5.5. To perform token-level alignments, we use MultiMirror (Procopio et al., 2021).[6]

4.2 Datasets

We experiment on three distinct language pairs: EN→DE, EN→ES and EN→FR. Following (Emelin et al., 2020), we gather the data from WMT14 for German and French and WMT13 for Spanish, considering only sentences coming from either CommonCrawl, News Commentary or Europarl, to maintain similar order of magnitudes among language pairs (and to contain pre-processing and training times). As validation sets, we employ *newstest2014* for EN→DE, *newstest2013* for EN→FR and *newstest2012* for EN→ES. All datasets employed in this work are freely available for research purposes.

Sense-Enhanced Datasets

We process each parallel sentence of the considered corpora with the procedure described in § 3.2, taking into account only content words whose Part-of-Speech tag is noun, as the challenge sets we evaluate upon only target nominal words.[7]

For POS-tagging and lemmatization we use Stanza (Qi et al., 2020). As disambiguation system, we use EWISER (Bevilacqua and Navigli, 2020), a neural WSD model based on BERT (Devlin et al., 2019), which has attained state-of-the-art performances on English as well as other languages. EWISER has been trained on SemCor (Miller et al., 1993) – the standard training set for WSD – and the WordNet Gloss corpus (Langone et al., 2004) – a semi-automatically annotated dataset featuring sense definitions. Detailed statistics of the base and parallel corpora produced are provided in § A.5.

Translation Test Set

We evaluate standard translation quality through the *newstest* datasets available in the specific WMT year (i.e., WMTXX corresponds to *newstest20XX*). The standard evaluation is carried out by means of SacreBLEU (Post, 2018), with signature `BLEU+case.mixed+numrefs.1 +smooth.exp+tok.13a+version.1.5.1`.

Disambiguation Bias Challenge Sets

To measure the disambiguation bias of each model we employ the challenge sets introduced by Emelin et al. (2020), composed of sentences reserved from the WMT14 English→German corpus. These challenge sets are based on sense clusters built by automatically merging together BabelNet synsets, which then are manually refined to ensure their correctness. Each sense cluster contains an English polysemous word and a set of German monosemous terms, which uniquely identify a certain meaning.

These clusters are used to create the following two challenge sets: WSD Bias and Adversarial. The former quantifies the intrinsic bias the model learned during training, while the latter measures how sensitive the model is to the insertion of terms that are usually associated with another sense cluster during training. Both challenge sets evaluate in terms of accuracy of correct disambiguation. A more detailed description of these datasets and their evaluation process is provided in § A.1.

DiBiMT We also evaluate on the German and Spanish portions of DiBiMT (Campolungo et al., 2022), a recent fully-manually annotated disambiguation bias challenge set, where models are asked to translate English sentences containing ambiguous words, and their translations are checked for either correct or incorrect translation equivalents, which, in contrast to previous benchmarks, are annotated manually and depend on the context of the sentence instead of relying solely on the sense of the source word.

4.3 Comparison Systems

We compare our sense-enhanced model with the following architectures:

[6] With a fallback strategy to fast-align (Dyer et al., 2013) in case no alignment is produced.

[7] We filter out all nouns appearing in the stopwords list provided by BabelNet.

1. *OPUS* (Tiedemann and Thottingal, 2020): a strong bilingual model which uses the same architecture and parameter count as ours, although it was trained on order of magnitudes more data;

2. *MBart-50* (Tang et al., 2021): the English-to-many version of the MBart-50 model;

3. *Baseline*: our base NMT models, trained on the datasets described in § 4.2.

In what follows we refer to our model fine-tuned with SCR as *Baseline+SCR*.

We note that, due to the way in which the WSD Bias Challenge Sets were constructed (i.e., by using sentences reserved from WMT14, see § 4.2), any fair evaluation against OPUS and MBart-50 is to be considered impossible, as such models have seen the sentences in the challenge sets during training. We therefore evaluate these two models only on standard BLEU, and point out that the resulting scores should only be regarded as references for our models' competence in the translation task.

5 Results

In what follows, first, we show that our model attains BLEU scores in the same ballpark as state-of-the-art approaches such as OPUS and MBart-50, despite the large gap in terms of parameters or training data. Then, we focus our evaluation on the WSD Bias, and compare our full-fledged model (*Baseline+SCR*) against its baseline variant.

5.1 General Translation Quality

In Table 1 we observe that the trained baselines are more than competent in the translation task: indeed, when considering average BLEU scores, they place between OPUS, which is trained on much more data but has the same parameter count, and MBart-50 (Tang et al., 2021), which is ~8 times larger but is capable of translating English to 50 languages.

In contrast to common debiasing techniques, which often observe a degradation in performance on standard benchmarks (Clark et al., 2019; He et al., 2019), we report consistent BLEU improvements on all language pairs, all of which are statistically significant at different p-values (Table 1), providing empirical proof that the proposed method does not hurt the model's general translation capability, while at the same time it helps models generate less biased translations (as will be discussed in the upcoming sections).

5.2 Disambiguation Bias

Results on the Disambiguation Bias Challenge Sets (§ 4.2) are reported in Table 2, for both of which we show improvements: on the WSD Bias Challenge Set, the bias is reduced, significantly, by more than 1%; similarly, on the Adversarial Challenge Set, we see a reduction of homographs mistakenly disambiguated due to the injection of adversarial adjectives of 0.27%. We attribute this lower impact to the artificial nature of the adversarial sentences, some of which, by manual inspection, display poor grammatical fluency.

5.3 WSD Performance

We conduct an analysis of the performance of EWISER on the English sentences of the WSD Bias Challenge Set, to see how it fares in comparison with our NMT models. Unfortunately, as the sense clusters are not directly associated with BabelNet synsets, we reconstruct this association automatically and manage to retrieve only 1847 of the 3000 sentences in the challenge set.

Having retrieved BabelNet synsets for the target terms, we can apply EWISER and check whether the disambiguated synset matches one of the synsets retrieved for the sense cluster of the challenge sentence. Let us consider our running example, "The energy comes from a distant plant.", one last time: if EWISER disambiguates the term *plant* to its sense of *organism*, we count it as a mistake, similarly to the case where our NMT model translates it as *pianta* instead of *impianto* (i.e., its sense of *factory*). With this in mind, we evaluate EWISER, Baseline and Baseline+SCR on the aforementioned subset of sentences; we report the results of this evaluation in Table 2 (bottom).

The results indicate that, for this setting, both NMT models actually perform quite a lot better than a pre-trained disambiguation system. One reason for this might be the different distributions the models are trained on: by design, the challenge sentences follow a distribution similar to the corpus used to train the NMT model, whereas EWISER is trained on sentences coming from news corpora from the 1960s and dictionary-like definitions. Moreover, in theory, if we were to apply the refinement process described in § 3.2 to disambiguate the challenge sentences, we would achieve a perfect score, as the target German lemmas are monosemous and thus the disambiguation is im-

[8] $10k$ bootstrap samples of 50% the test set's size each.

	EN → DE		EN → FR	EN → ES
	WMT14	WMT19	WMT14	WMT13
OPUS†	27.58	39.39	39.93	35.00
MBart-50‡	25.60	35.80	36.12	29.50
Baseline	26.34	36.93	38.05	32.82
Baseline+SCR	**27.26**	**37.74**	**38.48**	**33.18**
Baseline+SCR$_{-KL}$	26.13	36.45	37.85	33.15
Baseline+SCR$_{-ARES}$	25.75	35.93	37.33	32.49
Baseline+SCR$_{-AR}$	26.11	36.74	37.38	32.93
Baseline+SCR$_{\text{RAND}}$	25.63	34.79	/	/

Table 1: Standard evaluation results. Numbers represent SacreBLEU scores. Statistical significance is computed according to Paired Bootstrap Resampling (Koehn, 2004) w.r.t. the row above. Underlined numbers represent $p < 0.02$ and $p < 0.001$.[8] † represents systems that accessed more training data than us, but with the same parameter count. ‡ represents systems that, beside access to a larger pool of data, also feature bigger underlying models.

plicitly solved. The results of using EWISER's raw annotations are discussed in § 5.5.

Finally, we choose not to perform a similar comparison on the Adversarial Challenge Set, as its examples are designed to specifically target NMT models via adversarial injections; we leave studying their impact in WSD systems as future work.

MODEL	WSD Bias ↓	Adversarial ↓
Baseline	12.27	4.48
Baseline+SCR	**11.23**	**4.21**
Baseline+SCR$_{-KL}$	12.43	5.14
Baseline+SCR$_{-ARES}$	12.53	4.75
Baseline+SCR$_{-AR}$	13.07	4.93
Baseline+SCR$_{\text{RAND}}$	12.56	5.04
EWISER	13.70	/
Baseline$_{cf}$	11.86	/
Baseline+SCR$_{cf}$	9.91	/

Table 2: Results on WSD Bias Challenge Sets. Numbers represent error rates (lower is better). Underlined results represent statistical significance at $p < 0.001$, compared to the row above, according to McNemar's test (McNemar, 1947).

5.4 System Examples

In Table 3, we report some examples of disambiguation corrected by our model according to the WSD Bias Challenge Set. The baseline is translating the terms to their most frequent sense (column *Wrong sense*), instead of the correct one (column *Target sense*). Moreover, the third example shows that this is not only a word matching task, as the improved model is able choose the correct subword and can capture the nuances of meaning in more uncommon senses.

5.5 Ablation Study

Ablation on SCR To measure the importance of the KL term in the loss, we fine-tune the model without including it in the SCR objective (§ 3.3) and report the results in Tables 1 and 2 (row Baseline+SCR$_{-KL}$). We observe that, without KL, the model struggles to leverage the double inputs efficiently; indeed, its translation performance drops around 1 BLEU point on average, while the error rates increase by roughly 1% on both bias challenge sets. These results back our intuition that the KL divergence helps to distill sense information from the sense-enhanced inputs, and is indeed a crucial component to our formulation.

Ablation on ARES We also test our system replacing the pre-trained sense embeddings provided by ARES with randomly initialized learnable embeddings and report this result in Tables 1 and 2 (row Baseline+SCR$_{-ARES}$). As expected, both translation quality and disambiguation bias drop consistently. Indeed, learning sense embeddings from scratch is much harder than learning a mapping between a fixed space and a trainable one.

Ablation on Annotation Refinement We evaluate our sense Annotation Refinement process (§ 3.2) by fine-tuning the model on the unconstrained sense annotations provided by EWISER (Baseline+SCR$_{-AR}$), i.e., by considering the synset with the highest confidence on the source word as the correct one, instead of S^*. In the bias

Source sentence / Reference sentence / Baseline output / Enhanced output	Target sense	Wrong sense
S: [...] that both first words start with the same **letter**. **R**: [...] dass beide Begriffe mit demselben **Buchstaben** beginnen. **B**: [...] dass beide Wörter mit dem gleichen **Brief** beginnen. **E**: [...] dass beide Wörter mit dem gleichen **Buchstaben** beginnen.	*alphabet symbol*	*written message*
S: At least since the **fall** of 2008, leading economies' officials have agreed [...] **R**: Spätestens seit **Herbst** 2008 stimmen die Vertreter führender [...] **B**: Zumindest seit dem **Fall** 2008 haben sich die Beamten [...] **E**: Zumindest seit dem **Herbst** 2008 haben sich die Beamten [...]	*season*	*act of falling*
S: The construction of the Deurganck dock **lock** is [...] **R**: Der Bau der **Schleuse** am Deurganck-Dock ist [...] **B**: der Bau der Deurganck-**Hafensperre** ist [...] **E**: der Bau der Deurganck-**Hafenschleuse** ist [...]	*segment of a canal*	*blockade*

Table 3: Examples of sentences that were disambiguated correctly by our enhanced model but not by the baseline. Ambiguous word is in **blue**, wrong translation is in **red**, correct translation is in **green**.

Model	EN → DE	EN → ES
OPUS†	27.99	36.66
MBart-50‡	28.73	33.89
Baseline	24.00	26.44
Baseline+SCR	25.00	25.84

Table 4: Accuracy scores on DIBIMT. † and ‡ have the same meaning as in Table 1. Higher is better.

evaluation (Table 2), the performances on both challenge sets drop significantly ($p < 0.001$), which is in line with EWISER's performance on this challenge set (§ 5.3). Furthermore, the BLEU scores drop too, although not as significantly (Table 1), but still always under-performing with respect to Baseline+SCR.

Ablation on Sense Annotations Finally, we test whether the sense annotations have an impact by replacing them with random senses for the specific word, drawn from the sense vocabulary with uniform probability, during the fine-tuning stage (Baseline+SCR$_{\text{RAND}}$).[9] As expected, we observe that randomly injecting senses is detrimental, with important performance drops in both the standard and the bias evaluation benchmarks.

5.6 Evaluation on DIBIMT

In Table 4 we report the results obtained on DIBIMT (Campolungo et al., 2022). For the sake of conciseness, we only report accuracy scores as a proxy for the general disambiguation bias displayed by our models.

While on English→German we observe an improvement of 1%, the performance on English→Spanish decreases by around 0.6%. We hypothesize that our English→Spanish model might be undertrained, as its accuracy differs by around 10% from OPUS, its direct comparison, while on English→German the difference is only of around 3%. We leave further investigation of this issue, including training larger, more capable models, as future work.

6 Conclusions

In this paper, we presented a fine-tuning strategy that, by leveraging the explicit sense annotations produced by a novel high-precision technique, effectively reduces the disambiguation bias of a baseline Neural Machine Translation model while at the same time also strengthening translation performances, without introducing any requirement at inference time.

Our analysis on a strong disambiguation system showed that its ability to disambiguate polysemous nouns is worse than that of a baseline NMT model, at least in the studied out-of-domain setting.

We believe that this work paves the way for better bias reduction techniques in MT, while also fostering interest in the issue represented by the disambiguation bias. As future work, we plan to further study the ability of NMT models to perform Word Sense Disambiguation and to strengthen research at the intersection of these two fields, with a view to building stronger and more reliable models.

[9] Due to time constraints, we only perform this ablation on the English→German model.

Acknowledgements

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the PerLIR project (Personal Linguistic resources in Information Retrieval) funded by the MIUR Progetti di ricerca di Rilevante Interesse Nazionale programme (PRIN 2017).

This work was also partially supported by the MIUR under the grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagatioN for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844. International Joint Conferences on Artificial Intelligence Organization.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 44–49, Bucharest, Romania. Global Wordnet Association.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online and Dublin, Ireland. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan. Association for Computational Linguistics.

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)*, pages 61–72, Prague, Czech Republic. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, Vancouver, Canada. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Georgia. Association for Computational Linguistics.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2019. Widening the representation bottleneck in neural machine translation with lexical shortcuts. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 102–115, Florence, Italy. Association for Computational Linguistics.

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

Viktor Hangya, Qianchu Liu, Dario Stojanovski, Alexander Fraser, and Anna Korhonen. 2021. Improving machine translation of rare and unseen word senses. In *Proceedings of the Sixth Conference on Machine Translation*, pages 614–624, Online. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 63–69, Boston, Massachusetts, USA. Association for Computational Linguistics.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs 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 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *Int. J. Lexicogr*, 3(4):235–244.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proc. of the Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Quang-Phuoc Nguyen, Anh-Dung Vo, Joon-Choul Shin, and Cheol-Young Ock. 2018. Effect of word sense disambiguation on neural machine translation: A case study in korean. *IEEE Access*, 6:38512–38523.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 2021. MultiMirror: Neural cross-lingual word alignment for multilingual word sense disambiguation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3915–3921. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018a. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018b. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5905–5911, Marseille, France. European Language Resources Association.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, Online. Association for Computational Linguistics.

Zikang Wang, Linjing Li, and Daniel Zeng. 2020. Knowledge-enhanced natural language inference based on knowledge graphs. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6498–6508, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Warren Weaver. 1949. Translation. *Machine Translation of Languages: Fourteen Essays*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Bias Evaluation Challenge Sets

We here provide a more detailed description of the datasets introduced by (Emelin et al., 2020). From § 4.2, recall that these challenge sets are based on sense clusters built on BabelNet, where each sense cluster contains an English polysemous word and a set of German monosemous terms, which uniquely identify a certain meaning.

We highlight that there is no direct link between the sense clusters and the data produced by our Annotation Refinement process, as the sense clusters are i) heavily manually refined[10] and ii) based on the entire BabelNet4 inventory (16M concepts), while EWISER only covers the subgraph of BabelNet linked to WordNet (117k concepts), as is common in the multilingual WSD setting. As such, we do not consider the evaluation to be in any way more favorable towards our system.

WSD Bias contains sentences whose targeted English term is likely to be translated into a specific different sense due to co-occurrences of words in the sentence itself. For example, in the sentence "a lot of money was spent to renovate the *capital*" the word *capital* is likely to be translated into its sense of *amount of money* due to the presence of the words *money* and *spent*. A mistake is detected if the term is translated into any of the German words contained in the most likely sense cluster. The goal of this task is to measure the intrinsic bias the model learned during training.

Adversarial contains two sets of sentences, the original sentence and its adversarial counterpart, built by injecting an adjective that is likely to flip the disambiguation performed by the NMT model towards a specific sense. For example, given the sentence "they met in the spring of 2020", the adversarial example would be "they met in the *hot* spring of 2020". The injection of *hot* leads the model to translate *spring* into its sense of *water source* as opposed to its correct sense of *season*. A mistake is detected every time the non-adversarial sentence is translated into the correct sense, whereas its adversarial counterpart is flipped to the sense cluster the adjective points to. The goal of this task is to measure how sensitive the model is to the insertion of terms that are usually associated with another sense cluster during training.

[10] As discussed in § 5.3, almost 40% of the challenge instances could not be linked back to BabelNet synsets, further confirming the impact of the manual refinement performed.

A.2 Training a Sense-Enhanced NMT Model

Our work is based on the assumption that providing a neural model with sense annotations for ambiguous words helps in disambiguating them. While this is rather intuitive, and has been shown to be the case in previous works (Nguyen et al., 2018; Pu et al., 2018b), we test this hypothesis in our setting by training an NMT model, from scratch, with sense-enhanced sentences only (see § 3.3 for details). We train a model comparable with the Baseline (i.e., same architecture and hyperparameters) on the English→German training set (§ 4.2), and observe that it achieves higher BLEU scores than the Baseline (which is trained on the same data but with plain sentences). For instance, the sense-enhanced model achieves a BLEU score of 27.22 on WMT14 and 36.79 on WMT19, with the first being a statistically significant improvement. This confirms, once again, that sense-enhanced NMT models are on par or better than plain NMT models, although they introduce the heavy requirement of WSD at inference time, which our work aims at dropping.

A.3 Reproducibility Details

Preprocessing Times The preprocessing of the datasets needed to apply Annotation Refinement (lemmatization, Part-of-Speech tagging and then disambiguation through EWISER) required around 4 days in total on an RTX 2080 Ti (roughly 3M sentences per day).

Training infrastructure and duration All our experiments were carried out on either an NVIDIA RTX 2080 Ti or a RTX 3090, depending on availability.

Model training required on average 4 days on a 3090, 7 days on a 2080 Ti. Fine-tuning epochs required around 10 hours each (on a 3090), with most finishing due to early stopping before the end of the second epoch.

Parameter counts We used HelsinkiNLP MarianMT models available on HuggingFace Transformers (Wolf et al., 2020) (e.g., for EN→DE, the model name is `Helsinki-NLP/opus-mt-en-de`). For

MODEL	WSD Bias		Adversarial		
	Correct ↑	%Error ↓	Correct ↑	%Error$_{ATTR}$ ↓	%Error$_{OTH}$ ↓
Baseline	71.37	12.27	86.10	4.48	0.40
Baseline+SCR	**73.27**	**11.23**	**87.36**	**4.21**	**0.34**
Baseline+SCR$_{-KL}$	70.37	12.43	85.30	5.13	0.40
Baseline+SCR$_{-ARES}$	70.53	12.53	85.75	4.75	0.45
Baseline+SCR$_{-AR}$	70.20	13.07	86.40	4.93	0.35
Baseline+SCR$_{\text{RAND}}$	68.83	12.56	84.51	5.04	0.63
EWISER	68.54	13.70	/	/	/
Baseline$_{cf}$	72.77	11.86	/	/	/
Baseline+SCR$_{cf}$	75.58	9.91	/	/	/

Table 5: Full results on WSD Bias Challenge Sets. Numbers represent percentages.

instance, EN→DE has 74.4M parameters, EN→ES has 77.9M, EN→FR has 75.1M.

For the fine-tuning stage we added ARES (frozen), thus adding a number of parameters equal to ARES's size (1536) times the number of unique synsets in the dataset (refer to Table 6 for approximate numbers). We also added a trainable projection layer of size $1536 * 512$ (512 is the Transformer's hidden dimension), thus adding $786k$ trainable parameters (which we drop after the fine-tuning).

Model training hyperparameters Similarly to (Emelin et al., 2020), we trained it on the entire dataset for a max of 100,000 steps with approximately 24k tokens per batch, label smoothing at 0.1 and an inverse square root learning rate scheduler with 4000 warmup steps. As optimizer, we used Adam (Kingma and Ba, 2015) with betas $(0.99, 0.98)$ and learning rate $7 \cdot 10^{-4}$, additionally employing an early stopping strategy with patience 5, monitoring the BLEU score on a validation set. We produced translations at inference time using a beam size of 5.

Fine-tuning hyperparameters For the fine-tuning, we resumed training using the weights of the baseline models, changed the learning to $1 \cdot 10^{-5}$ and reduced the warmup to 1000 steps; additionally, we evaluated the model every 10% of the fine-tuning steps rather than after each epoch, as we observed fast convergence during fine-tuning and multiple epochs were superfluous.

A.4 Disambiguation Bias Results

Table 5 reports the same results displayed in the paper, but includes the percentage of Correct translations for both challenge sets as well as the percentage of errors made from sentences that, after the injection of the adversarial adjectives, were translated into a sense that was neither the correct one, nor the one targeted by the adversarial injection (i.e., other).

A.5 Data Statistics

CORPUS	EN-DE	EN-ES	EN-FR
# sentences	4.13M	3.54M	5.09M
# tokens (src / tgt)	99.7M / 96.8M	94.7M / 98.7M	133M / 142M
# annotated sentences	2.97M	3.11M	4.25M
# annotations	6.5M	11.2M	13.6M
# EN terms vocab	634k	592k	808k
# EN terms covered	34.0k	40.0k	37.5k
# unique synsets	16.0k	20.5k	15.7k

Table 6: Training and fine-tuning produced data statistics.

A.6 Limitations of this work

Our work focuses on reducing the disambiguation biases picked up by NMT models during training. We acknowledge some limitations in our work:

1. Due to limited computational budget and the large number of resources required to train and fine-tune NMT models from scratch, we had to limit ourselves to one run per experiment, though, despite this, the consistency across languages seems to point to the empirical correctness of the claims.

2. We evaluated the bias reduction explicitly only on the English→German language pair. The reason for this was twofold: first, the datasets introduced by Emelin et al. (2020) only cover said pair, and require the accompanying training data be used in order to fully exploit the co-occurrences (and hence the biases) that the model is evaluated upon; second, upon manual inspection, we found that MuCoW (Raganato et al., 2019) contains many irrelevant candidates in its translation suite, and is in general very strongly affected by the noisy nature of BabelNet.

3. Our pipeline is strictly tied to both the accuracy of the multilingual WSD system employed and by the coverage of the underlying sense inventory. While EWISER and BabelNet work reasonably well for high-resource languages, the quality of the annotated corpus might decrease for low-resource ones.