Translate to Disambiguate: Zero-shot Multilingual Word Sense Disambiguation with Pretrained Language Models

Haoqiang Kang* **Terra Blevins*** **Luke Zettlemoyer**
Paul G. Allen School of Computer Science & Engineering,
University of Washington
{haoqik, blvns, lsz}@cs.washington.edu

Abstract

Pretrained Language Models (PLMs) learn rich cross-lingual knowledge and perform well on diverse tasks such as translation and multilingual word sense disambiguation (WSD) when finetuned. However, they often struggle at disambiguating word sense in a zero-shot setting. To better understand this contrast, we present a new study investigating how well PLMs capture cross-lingual word sense with Contextual Word-Level Translation (C-WLT), an extension of word-level translation that prompts the model to translate a given word in context. We find that as the model size increases, PLMs encode more cross-lingual word sense knowledge and better use context to improve WLT performance. Building on C-WLT, we introduce a zero-shot prompting approach for WSD, tested on 18 languages from the XL-WSD dataset. Our method outperforms fully supervised baselines on recall for many evaluation languages without additional training or finetuning. This study presents a first step towards understanding how to best leverage the cross-lingual knowledge inside PLMs for robust zero-shot reasoning in any language.

1 Introduction

Pretrained Language Models (PLMs) perform many cross-lingual tasks without explicit cross-lingual training signal, including word-level translation (WLT) across languages (Gonen et al., 2020). These models also demonstrate cross-lingual knowledge when finetuned for the word sense disambiguation (WSD) (Raganato et al., 2020; Pasini et al., 2021). However, the extent to which word sense knowledge comes from pretraining rather than finetuning is unclear: many PLMs struggle to disambiguate word sense when formulated as a binary classification task, the most common word sense setup for prompting language models (Shi et al., 2022; Scao et al., 2022).

To investigate this, we measure the ability of multilingual autoregressive language models to understand the cross-lingual meaning of words in a given context. Specifically, we extend the WLT task setup to include a specific context in the prompt, which we call Contextual Word-Level Translation (C-WLT). We empirically show that pretrained language models leverage contextual information in the prompt to improve WLT performance. In addition, both English and multilingual PLMs perform better on the contextual WLT tasks as model size increases, demonstrating improved cross-lingual knowledge at scale.

Translations of a word that change based on context are frequently due to differing word senses not shared by an analogous word in the target language (Resnik and Yarowsky, 1999). Inspired by this, we apply C-WLT to the task of WSD by translating the ambiguous word w in context with WLT and then assigning w with the senses in the overlap of the translated word's sense set with w's senses (Figure 4, left). We test this zero-shot approach for WSD on 18 languages from the XL-WSD dataset (Pasini et al., 2021). In our best setting, zero-shot WSD via C-WLT prompting outperforms prior supervised works on recall for many evaluation languages, even though our method requires no additional training on labeled WSD data. We also observe that ensembling diverse target languages with this method narrows down the predicted set of senses, as demonstrated by the improvements in Jaccard similarity with the reference set. Finally, we analyze our design choices and the types of errors made by this approach to better understand the behavior of WSD via C-WLT and how it relates to supervised WSD classification.

The overall findings of this work are as follows:

- PLMs leverage contextual information to encode cross-lingual knowledge and better capture lexical information, such as word translations and meanings.

*These authors contributed equally to this work.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1562–1575

- We can leverage this contextual knowledge of lexical translation to effectively perform *zero-shot* WSD for many languages, including low-resource ones and languages the PLM was not explicitly pretrained on.
- The efficacy of WSD via C-WLT depends on different factors such as pretraining languages, model size, and target language choice: smaller multilingual PLMs perform well on seen languages, but they are more sensitive to design choices and do not generalize as well as larger English PLMs.

In sum, we evaluate the lexical translation skills of PLMs in context, and we present a first step towards applying that skill to the downstream task of WSD. Given that most WSD training data outside of English are automatically created (e.g., Scarlini et al., 2019; Barba et al., 2021), and that annotating gold data incurs significant costs for each new language, zero-shot approaches such as our proposed WSD via C-WLT approach are crucial for improving WSD in low-resource languages.

2 Contextual Word-Level Translation

A standard method of evaluating the cross-lingual capabilities of PLMs is the task of a word-level translation (WLT), where the model is prompted to translate a word w_s from a source language L_s into another target language L_t (Gonen et al., 2020). However, this setup does not consider variations in the translation of w_s into L_t that occur when the surface form of w_s represents multiple meanings (i.e., senses) in different contexts.

We propose an extension of the word-level translation task, Contextual Word-Level Translation (C-WLT), which requires translating words correctly based on how they are used in a given context (Figure 4, right panel). Specifically, we prompt the PLM to translate w_s from L_s into L_t when conditioned on a specific context c_s where $w_s \in c_s$; we then measure whether it produced the correct translation(s) w_t in context of w_s.

For example, if we want to translate "plant" into Chinese based on the context sentence "The plant sprouted a new leaf", we prompt the PLM with *In the sentence "The plant sprouted a new leaf", the word "plant" is translated into Chinese as* __. This evaluation allows us to quantify a PLM's ability to align meaning across languages in a context-specific manner.

2.1 Experimental Setup

Prompts and Languages After a preliminary analysis of potential prompt formats, our experiments use the following prompts:

- **Without Context**: The word "w_s" is translated into L_t as __
- **With Context**: In the sentence "c_s", the word "w_s" is translated into L_t as __

We perform experiments with English as the source language and translate into Chinese, French, and Spanish as the target languages.

Models We use the GPT-Neo models (Gao et al., 2020) with sizes between 125 million to 20 billion parameters (including the GPT-J model that contains 6B parameters; Wang and Komatsuzaki, 2021) and the BLOOM series with different model sizes from 560 million to 7.1 billion (Scao et al., 2022). We note that BLOOM is explicitly pretrained on all three of our target languages, whereas GPT-NeoX (Black et al., 2022) is trained as an English LM; however, GPT-NeoX's pretraining corpus contains an estimated $\sim 2.6\%$ of non-English text (Gao et al., 2020), and prior work has found even small percentages of non-English text can facilitate cross-lingual transfer in English PLMs (Blevins and Zettlemoyer, 2022).

Dataset We first select candidate source words from the English inventory in the XL-WSD dataset (Pasini et al., 2021). We then create language pair datasets with <*source word, source example context, translations in context*> tuples, where the sense-specific translations and example contexts are obtained from WordNet (Miller, 1995). We filter these datasets to include examples where two senses (the most common sense and at least one other sense) meet the following criteria: (a) both senses have non-overlapping sets of translations in the target language, and (b) both senses are annotated with example contexts in the source language.

For each example, we use the target language translations of the paired, incorrect sense in that setting and 50 randomly selected words in the target language as incorrect translations as negative samples. Due to limited cross-lingual coverage with WordNet, the EN-FR, EN-ES, and EN-ZH experiments include 2448, 2470, and 2084 evaluation examples, respectively.

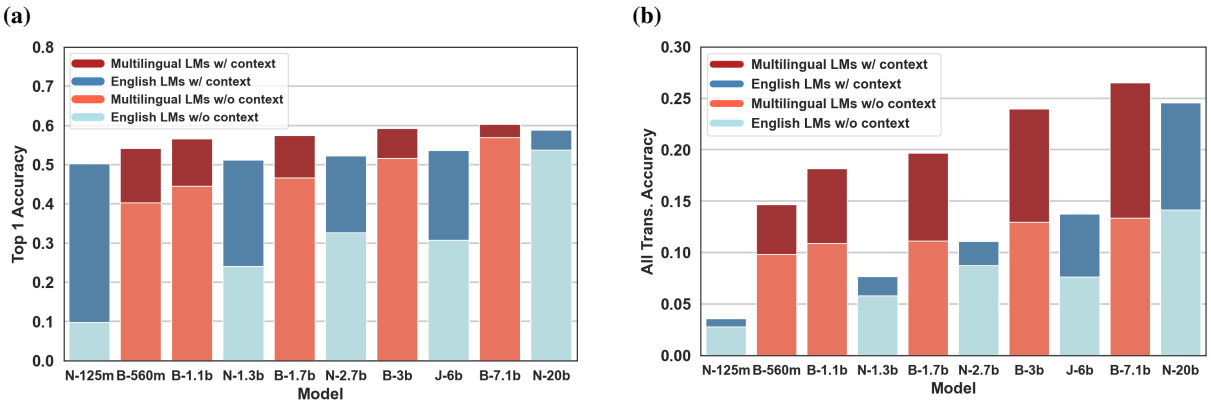

Figure 1: Results of the zero-shot contextual WLT accuracies on GPT and BLOOM family models of different sizes (a) The results of top-1 accuracies across models. (b) The results of all translations accuracies across models. N: GPT-Neo, B: BLOOM, J: GPT-J

Metrics We present three different metrics to evaluate models' performance on the WLT task, with and without context.

- **Accuracy**: We use two metrics to measure the models' accuracy. (1) *top-1 accuracy* measures the percentage of test instances in which the translation with the highest log-likelihood is one of the correct translations for a given sense. (2) *All translations accuracy* measures the percentage of test instances where all k correct translations for that sense are assigned the k highest likelihoods by the model.

- **Negative Log-Likelihoods (NLL)**: We compare the average *negative log-likelihood (NLL)* of all (1) correct and (2) incorrect translations for each sense, as well as (3) the *ratio* of the average NLL of the top-1 correct translations to the average NLL of all incorrect translations for each sense.

- **Error Reduction**: We evaluate the impact of adding context sentences on resolving two types of errors. The first is *disambiguation* errors, where the model produces a valid translation without context that would be incorrect in the additional context; the second is *translation* errors, where the model correctly translates the word in question (based on the context sentence) but produces a mistranslation without context.

2.2 Results

Adding Context Improves Word-Level Translation Accuracy Figure 1 presents the overall WLT results with and without context, averaged across the three target languages; word-level translation performance improves across all settings with the addition of context.[1] We also observe that the performance of both uncontextualized and contextualized word-level translation improves as the model size increases, which corroborates prior findings that larger models better capture cross-lingual information from pretraining (e.g. Lin et al., 2022).

Our experiments also show that, on average, the multilingual models outperform comparably sized English models in both WLT settings: the multilingual models achieve an average *top-1 accuracy* of 47.94% in the uncontextualized task and 57.51% in the contextual task, whereas the English models obtain 30.20% and 53.2% in these settings, respectively. However, the performance gap between English and multilingual models narrows when we add sentences that use the word in context. Specifically, the experiments show that the largest English model, GPT-NeoX, performs similarly to the (smaller) multilingual BLOOM models; this suggests that English language models become more effective in leveraging limited cross-lingual knowledge at larger scales.

While these trends are generally consistent across languages, we observe some variation (Appendix D). For instance, smaller English models perform notably worse on EN-ZH than when translating into FR and ES, likely because it is more difficult to generalize to languages written in a different

[1]The results for individual target languages can be found in the appendix. (Figure 7 for Chinese; Figure 8 for French; Figure 9 for Spanish)

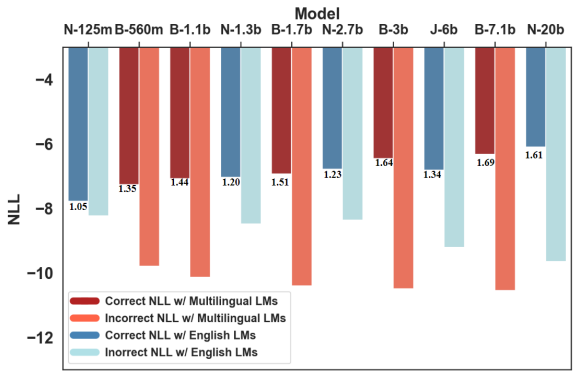

Figure 2: The average NLL of all correct and incorrect words across models in the contextual WLT analysis (less negative is better). Numbers represent the NLL ratio of incorrect to correct translations.

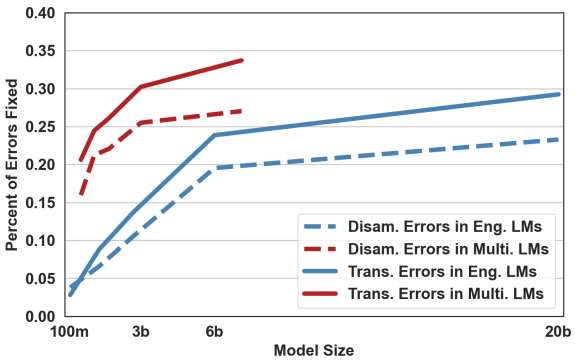

Figure 3: The impact of adding context to WLT on translation (trans.) and disambiguation (disam.) errors.

script (Blevins and Zettlemoyer, 2022). Furthermore, English models generally perform similarly to multilingual models on EN-ES translation.

Finally, in the setting of *all translations*, we observe that performance improvements with the addition of context are more significant for multilingual models than for English ones, leading to larger performance gaps between these types of models in the C-WLT setting.

Negative Log-Likelihoods We also consider the negative log-likelihoods of each model for the top correct translation compared to incorrect translations (Figure 2). These results show that the correct translations' negative log-likelihood (NLL) improves as the model size increases, suggesting that the models become more confident in their predictions in absolute terms. Furthermore, we find that the NLL ratio between correct and incorrect translation words generally increases as the model size improves; the multilingual models also demonstrate better differentiation ability between correct and incorrect translations than English models. Specifically, we observe an average ratio of 1.53 between incorrect and correct translations for multilingual models, compared to 1.28 for English models.

Translation Error Reduction with Context Finally, we analyze the extent to which adding context sentences resolves errors made by the PLMs in the standard WLT setting (Figure 3). Our results show that larger models benefit more than smaller ones from using contextual information to correct translation errors, with a greater percentage of prior errors resolved with the addition of context; this further highlights their ability to leverage the additional context better. In addition, multilingual models fix errors at a higher rate than English models when given context.

Surprisingly, we also observe that context helps correct complete *translation* errors at higher rates than it does to *disambiguate* the appropriate translation given a context sentence. This behavior generally holds for both the English and multilingual models and across all model scales. The smallest English models are an exception where very few errors of either type are resolved by context, despite their overall performance significantly improving in the C-WLT setting.

3 Zero-shot Word Sense Disambiguation via C-WLT

Building on the intuition from the previous section that contextual word-level translation can differentiate between different meanings of a word in the source language, we apply C-WLT to the task of multilingual word sense disambiguation (Figure 4). Specifically, we propose a two-step process wherein we (1) prompt the PLM for C-WLT to translate the ambiguous target word, w, in the relevant context and (2) disambiguate w based on the senses of its translation.

For instance, to disambiguate the word "plant" as it is used in the context "The plant sprouted a new leaf", we first prompt the PLM to translate "plant" into the chosen target language (e.g., Chinese) with the C-WLT setup from the previous section. We then take the PLM's top translation (in this case, "植物") and obtain its senses from a multilingual word sense ontology. We then label the example with the senses shared by "plant" and "植物".

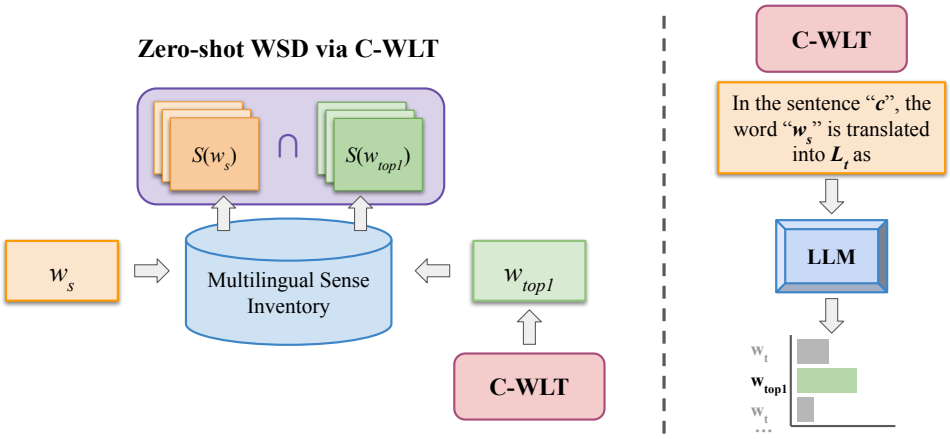

Figure 4: Overview of the proposed method for multilingual WSD via C-WLT (left) and the prompting setup for C-WLT (right). We translate each ambiguous word w_s in context into a target language t with a PLM and label it with the intersection of its labels and the labels of the translation w_{top1}.

3.1 Method

The goal of word sense disambiguation (WSD) is to determine the meaning of the word w in a specific context c and label it with the sense label (or labels) that represents this meaning out of the candidate set of senses associated with that word, S. In our proposed approach, **WSD via C-WLT**, w and c are in a language L_s, and word senses are from a multilingual ontology (BabelNet, Navigli and Ponzetto, 2010) and shared across languages.

First, we prompt a PLM with the C-WLT setting to translate w_s based on c_s into the target language L_t. We then obtain the inventory of all possible translations of w_s into L_t from the multilingual word sense ontology and rank them with the PLM conditioned on the C-WLT prompt. We then label w_s with the set of senses in the intersection of its candidate senses, $S(w_s)$, and those of the top-scoring translation under the PLM, $S(w_{top1})$. This means the WSD via C-WLT method assigns a set of labels to w rather than a single sense label, unlike most supervised WSD classifiers.

Ensembling Target Languages The described method for WSD via C-WLT obtains potential senses from translating into a single target language. We extend the method to ensemble the senses from a set of target languages T, as we hypothesize that senses shared by translations of w_s in multiple typologically diverse languages are more likely to be relevant to the specific context at hand. This is supported by Bao et al. (2021), which argues that every sense can be disambiguated with translation if *all* possible languages are considered.

Specifically, we consider the multiset of senses for the top translation in every target language: $S(T) = \{S(w^t_{top1}) : t \in T\}$. Our target set $S(T)'$ is the subset of $S(T)$ that contains all senses with the highest multiplicity (i.e., occur most frequently) in $S(T)$. This means that senses shared by translations of w_s in multiple languages are more likely to be included in $S(T)'$. Similar to the single target language setting, we obtain the final predicted sense set from the intersection of $S(T)'$ and $S(w_s)$.

3.2 Experimental Setup

Datasets We evaluate performance with the XL-WSD dataset (Pasini et al., 2021), which covers 18 languages: Basque, Bulgarian, Catalan, Chinese, Croatian, Danish, Dutch, English, Estonian, French, Galician, German, Hungarian, Italian, Japanese, Korean, Slovenian, and Spanish. We use the BabelNet (Navigli and Ponzetto, 2010) multilingual word sense ontology to obtain translations and sense inventories of the data.

We consider five target languages for our experiments: English, Chinese, Russian, Spanish, and Finnish. Our choice of target languages aims to cover semantically diverse target languages (to increase variety in the translation to sense mappings) while maintaining high coverage within the multilingual ontology.[2] When a (non-English) evaluation example does not have at least one corresponding translation in the target language, we back off to the English translation setting as it provides

[2]English covers 100.0% of the evaluation examples (excluding EN-coarse), while Chinese, Spanish, Finnish, and Russian cover 79.0%, 95.3%, 99.6%, and 60.0%, respectively.

Language	MCS	Prior Work*	Recall			Jaccard Index		
			NeoX	B-3B	B-7.1B	NeoX	B-3B	B-7.1B
Basque	32.72	51.71 (b)	47.85	52.53	**54.31**	37.20	41.04	**42.95**
Bulgarian	58.16	73.60 (c)	**75.51**	71.56	72.05	**66.28**	63.32	63.78
Catalan	27.17	**57.47** (b)	55.73	55.83	56.40	39.44	40.41	**40.85**
Chinese	29.62	57.05 (b)	**61.03**	60.64	58.87	**46.86**	46.78	46.26
Croatian	62.88	74.40 (b)	**77.01**	74.85	74.82	**70.00**	68.53	68.46
Danish	64.33	81.80 (c)	**81.86**	76.76	77.38	**73.50**	69.69	70.32
Dutch	44.61	61.95 (b)	**66.25**	61.89	63.46	**55.72**	52.07	53.33
English†	63.37	**80.40** (c)	72.61	72.15	73.20	60.56	60.13	**61.39**
Estonian	46.87	68.88 (b)	**70.24**	65.58	65.88	**61.72**	58.94	58.80
French	59.31	**83.88** (a)	76.04	76.47	78.02	64.67	65.62	**68.00**
Galician	60.85	67.30 (c)	74.15	74.63	**74.82**	60.47	**61.06**	60.84
German	75.99	**84.69** (b)	81.45	78.31	81.57	**74.40**	71.60	74.02
Hungarian	47.29	**76.40** (c)	75.52	71.56	72.04	**66.28**	63.32	63.77
Italian	52.77	**77.80** (c)	76.63	74.50	74.58	**57.91**	57.62	57.63
Japanese	48.71	67.47 (b)	**71.63**	70.78	71.38	57.56	57.38	55.72
Korean	52.48	**68.20** (c)	66.39	67.52	67.73	60.95	61.01	61.46
Slovenian	36.71	**68.36** (a)	53.12	46.21	47.93	**40.32**	33.36	37.05
Spanish	55.65	76.93 (b)	75.42	75.53	**77.66**	55.58	56.50	**58.36**
Avg.	49.31	–	70.35	68.62	69.45	58.59	57.42	58.24

Table 1: Zero-shot Recall and Jaccard Index for multilingual WSD on the XL-WSD dataset in the best-ensembled setting. Results for languages on which Bloom was pre-trained are underlined. *Prior work numbers are drawn from the best (fully-supervised) results reported in (a) Pasini et al. (2021), (b) Berend (2022), and (c) Zhang et al. (2022). †For the 1512 (out of 8062) English examples with coverage issues, we used MCS as predictions.

full coverage over all non-English evaluation sets. When evaluating English, we instead back off to the most common sense (MCS) of the word when the target language(s) does not cover an example in each evaluation setting.

Models Picking the three most powerful PLMs from the previous section, we use the BLOOM models with 3 billion parameters and 7.1 billion parameters and the GPT-NeoX model with 20 billion parameters. While GPT-NeoX is primarily trained on English, the Bloom models are specifically pretrained on 6 out of the 18 evaluation languages of the XL-WSD dataset (Basque, Catalan, Chinese, English, French, and Spanish).

Baselines We compare our approach with the Most Common Sense (MCS) baseline, which predicts each word's most common sense according to BabelNet (Pasini et al., 2021). We also report the best results from the models benchmarking XL-WSDin Pasini et al. (2021) as well as those in Zhang et al. (2022) and Berend (2022). We present prior results as a point of reference; however, these previous models for the XL-WSD dataset require supervised training with annotated WSD data, unlike our zero-shot approach, which assumes no additional data or finetuning of the PLM.

Evaluation Metrics for WSD via C-WLT We consider two automatic metrics for evaluating the performance of the WSD via C-WLT approach. The first is *recall*, or how often the predicted label set contains at least one of the gold annotations for a given example. This metric is obtained from the XL-WSD evaluation script and is the standard evaluation for this benchmark; it is often reported as (and is equivalent to) F1 or accuracy in cases where the WSD model produces a single prediction.

However, recall overestimates performance in cases where a WSD approach predicts many unrelated sense labels in addition to a correct one. Therefore, we also calculate the *Jaccard index* between the predicted set and the reference set of sense labels for each example: $\frac{|L_{true} \cap L_{pred}|}{|L_{true} \cup L_{pred}|}$. While the Jaccard index is a better automatic measure of similarity for sets than recall, the metric can underestimate performance in cases where other, closely related senses are appropriate in the given context yet not included in the reference sense set.[3]

We note that the Jaccard index is closely tied to F1 score: the two metrics are monotonically related and will give the same relative performance across methods. In terms of (true and false) positives and negatives, Jaccard Index is $\frac{\text{TP}}{\text{TP+FP+FN}}$, whereas F1 score is $\frac{\text{TP}}{\text{TP}+\frac{1}{2}(\text{FP+FN})}$. We report Jaccard index as it is an established metric for set similarity.

[3]This type of annotation error is the most common found in an audit of English WSD corpora (Maru et al., 2022).

Target Lang.	Recall	Jaccard	Delta*
Spanish	74.23	52.94	20.0
English	67.16	53.37	11.7
Finnish	66.35	54.28	12.9
Russian	67.42	55.08	10.2
Chinese	70.84	57.77	9.6
Best Setting	70.35	58.59	8.7
All 5 Joint†	66.60	57.50	6.7

Table 2: Average Recall and Jaccard Index for target language settings on the GPT-NeoX model, as well as the delta(*) increase in sense prediction rates. †"All 5 joint" uses all of the above target languages, whereas "best setting" ensembles English, Chinese, and Russian.

4 Multilingual WSD Results and Analysis

We first present the performance of our method for multilingual WSD on the two automatic metrics, recall and Jaccard index, and compare this approach to prior work on this task (Section 4.1). We then consider the effect of ablating different modeling choices on our method (such as the choice of target language for C-WLT and prompt language; Section 4.2), and we analyze the types of errors the approach produces more closely (Section 4.3).

4.1 Results

The multilingual WSD results are summarized in Table 1. In our experiments, we found that the best setting for achieving a balance between recall and Jaccard Index was to ensemble English, Chinese, and Russian as the target languages with English prompts (Table 2). The results show that our approach achieves higher recall than the prior works in 11 out of the 18 source languages, despite our method being performed zero-shot from a pretrained language model. Considering recall as an upper-bound measure of performance, this result shows that translation-based approaches for WSD identify correct sense label(s) as well as or better than supervised methods.

We also find that despite being primarily pretrained on English, GPT-NeoX (20B) achieves higher recall and Jaccard index scores than Bloom-7.1 on ten source languages; most settings where the multilingual model performs better are on its pretraining languages, with little generalization to other languages. Finally, despite the Jaccard index scoring lower (by definition) than recall, we see similar performance trends across languages and models between recall and the Jaccard index in this ensemble setting.

4.2 Modeling Ablations

Different Target Languages To investigate the effect of the target language(s) on contextual word-level translation in the WSD task, we consider five target languages: English, Chinese, Russian, Finnish, and Spanish. We also experiment with all combinations of these languages for the joint target language settings (Table 2).[4] We also calculate the *delta* increase in the sense prediction rates, normalized by the number of senses for each example, as a measure of how many more senses our method predicts over the supervised baselines. To obtain this *delta*, we compare the standard classification setting of predicting a single label per WSD example and the number of labels predicted by each target language setting: $\frac{1}{n}\sum_{i=0}^{n}\frac{|\hat{S}_i|}{|S_i|} - \frac{1}{n}\sum_{i=0}^{n}\frac{1}{|S_i|}$ where S_i is the candidate sense set for the ith evaluation example and $\hat{S}_i$ is the set of senses predicted by our approach.

Our ablations indicate a tradeoff between the Jaccard index and recall. For example, our approach achieves the highest recall performance using Spanish as the sole target language, but the resulting Jaccard index is worse than any other target setting we test. This behavior is likely because target languages more similar to the source (such as Spanish, which is closely related to many of the Western European source languages in the XL-WSD dataset) return a larger set of predicted senses, which in turn improves recall but at the expense of set similarity with the gold labels. This hypothesis is corroborated by the high delta increase of 20% in the predicted set size of the Spanish setting over the standard single-label predicted setting.

However, this undesirable behavior is mitigated when using dissimilar target languages to the source and ensembling diverse languages. In our best setting of ensembling English, Chinese, and Russian, we find that the delta increase in the predicted set size is only 6.7%, while the Jaccard index increases by ∼6 points over Spanish. Furthermore, this ensembled setting still often outperforms prior approaches on recall.

Prompts in Different Languages We then consider the effect of prompt language on the WSD via C-WLT method by ablating prompts in English, the evaluation source language, and the target language. The English, Chinese, French, and Span-

[4]We report the Bloom results in Table 6 in the appendix; we observe similar tradeoffs when using those models.

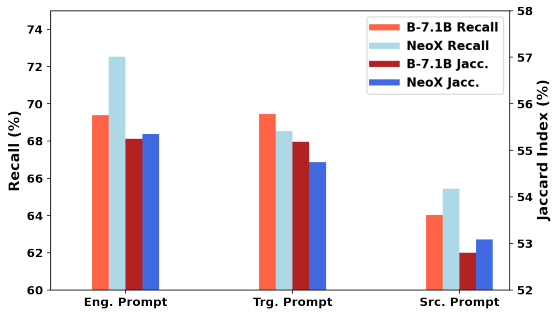

Figure 5: Effect of prompt language on performance.

Label Set	Recall		Jaccard	
	NeoX	B-7.1B	NeoX	B-7.1B
Orig.	63.78	57.74	52.01	50.98
Annot.	**74.01**	**74.54**	**54.29**	**52.73**

Table 3: Re-annotated and original label results on the re-annotated subset of the Chinese evaluation set.

ish prompts were obtained from or verified by native speakers; prompts in other languages are from Google Translate. We test two languages, Spanish and Chinese, as targets and evaluate (a) the overall performance of the method by the prompt language (Figure 5) and (b) the top-scoring prediction's language for each prompt setting, out of the union of the candidate word sets from the prompt, source, and target languages (Appendix Figure 6).

We observe that prompts in English and target languages outperform the source languages, with English prompts generally performing the best (though the target language prompts are comparable to English in Bloom). We also find that the non-English prompts are more likely to produce a top-1 prediction in the wrong (not target) language. This is particularly true in the case of source language prompts; the observed performance decrease suggests that prompting the model to generate a label in a different language than the prompt itself is difficult – unless the prompt language is English. Moreover, our results show that the multilingual LM (BLOOM-7.1b) is more prone to predicting words in the wrong languages than the English LM (GPT-NeoX).

4.3 Manual Precision Analysis

We observe that the gold annotations in the XL-WSD test sets mostly consist of one label. However, fine-grained word sense meanings are often similar or even overlapping, with fine-grained annotator agreement as low as 67% in some cases (Navigli, 2009). We hypothesize that other related senses may be suitable in many evaluation contexts but not included in the reference set.

To investigate this further, we ask three native language speakers to reannotate 392 examples of Chinese test data manually. This analysis finds that 172 examples (or 44%) have additional closely related senses not included in the original annotations. For example, consider the sentence: "广播还没说完，各班的同学早已纷纷冲出教室。"[5] In the XL-WSD dataset, the word "广播" is labeled with the definition, *"Be broadcast"*. However, our annotation adds a sense with the definition, *"Broadcast over the airwaves, as in radio or television"* into the reference set.

The results on the subset of the evaluation set show that, unsurprisingly, both models' recall and Jaccard index improve on the reannotated data (Table 3). We conclude that missing fine-grained annotations are one factor impacting our results. The many examples found during the analysis with other relevant senses indicate that the reference sets likely do not contain full coverage. This suggests that future research on multilingual WSD should consider the choice of reference sets to ensure that they reflect all relevant senses, as prior work has for English (Maru et al., 2022).

5 Discussion and Related Work

We first analyze the performance of PLMs in the new contextual word-level translation (C-WLT) setting to evaluate how well these models produce context-sensitive lexical translations. Other related work has instead tested the efficacy of prompting multilingual PLMs for sentence-level translation, such as Lin et al. (2022) and Vilar et al. (2022). Notably, Bawden and Yvon (2023) observe incorrect language prediction with multilingual PLMs, similar to our findings in Section 4.2.

We then apply the C-WLT setup to zero-shot multilingual WSD. This approach builds on Pasini et al. (2021), which highlights the role of multilingual language models in addressing the knowledge acquisition bottleneck problem in WSD. Other works have proposed different finetuning improvements to perform WSD better cross-lingually (Zhang et al., 2022; Berend, 2022). Unlike these approaches, our method does not require annotated training data, allowing it to generalize easily. Our

[5] In English, "Before the broadcast was finished, students from all classes had already rushed out of the classroom one."

proposed method is, to be best of our knowledge, the first attempt to apply large-scale autoregressive PLMs to word sense classification via in-context learning. Prior work on word sense prompting frames WSD as a binary classification task comparing a word's meaning in two contexts (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020).

More generally, WSD is closely related to and motivated by machine translation; Hauer and Kondrak (2023) outlines the relationship between lexical translation and WSD. A commonly proposed use case of WSD systems is to improve the translation of ambiguous words in MT; as such, multiple methods to incorporate word sense information (such as sense embeddings) into NMT systems have been proposed (e.g., Liu et al., 2018; Campolungo et al., 2022b). Furthermore, word sense knowledge has been used to evaluate NMT systems (Campolungo et al., 2022a). Prior work has also leveraged MT systems and data to improve an underlying WSD classifier (Luan et al., 2020) and automatically annotate WSD data (Diab and Resnik, 2002; Apidianaki and Gong, 2015; Hauer et al., 2021; Barba et al., 2021; Su et al., 2022). We build on this latter line of work's intuition to extrapolate word senses from the translations of ambiguous words in context.

6 Conclusion

In this work, we examine the ability of pretrained language models to utilize contextual information in cross-lingual settings. Specifically, we propose contextual word-level translation (C-WLT) and test different PLMs' ability to improve lexical translations in context. We then propose a zero-shot prompting technique for multilingual WSD, using C-WLT as a component. Our experiments show the method's effectiveness on 18 languages, including those not included in the PLM's pretraining.

The performance of WSD via C-WLT relies on the relationship between pretraining languages, model size, and the choice of the target language: smaller multilingual PLMs are more effective for languages on which they have been pretrained but are more sensitive to design choices, lacking the broad applicability of their larger English counterparts. Future research examining these interactions and their tradeoffs more closely is vital for improving zero-shot WSD approaches and building better cross-lingual applications of PLMs in general.

Limitations

We recognize several limitations that influence C-WLT and our proposed approach for WSD. First, the WSD via C-WLT method depends on the composition of the multilingual word sense ontology we use to obtain cross-lingual word senses and translations. Lower coverage in the chosen target language will hinder the method's performance: we see this empirically in the case of English as an evaluation language, as no target language setting (including ensembling) fully covers English, which requires us to back off the MCS of each word.

Similarly, the translation capability of PLMs, particularly for low-resource languages, may limit the effectiveness of both C-WLT and our WSD approach that relies on it. While we first present a study of the efficacy of C-WLT before incorporating it into our WSD method, due to data limitations (i.e., constructing a C-WLT data for each language pair that contains examples covering multiple senses of many different target words), we examine three high-resource language pairs. However, better cross-lingual PLMs can be directly integrated into our proposed approach as they are developed to improve multilingual WSD.

Finally, our approach is not well-suited for distinguishing between very fine-grained word senses. While our small-scale manual precision analysis (Section 4.3) suggests that at least some WSD evaluation sets are not annotated with complete coverage of all relevant senses – leading to an underestimate of our approach's performance – the ability to differentiate between closely related senses precisely remains a hurdle for the WSD via C-WLT method, and addressing this issue in the future will further improve its applicability.

Acknowledgements

We thank the human annotators for the manual precision analysis of the Chinese XL-WSD evaluation set. We also thank Hila Gonen for her helpful comments and discussion on this work.

References

Marianna Apidianaki and Li Gong. 2015. LIMSI: Translations as source of indirect supervision for multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 298–302, Denver, Colorado. Association for Computational Linguistics.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global WordNet Conference*, pages 1–7.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2021. Mulan: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3837–3844.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom.

Gábor Berend. 2022. Combating the curse of multilinguality in cross-lingual WSD by aligning sparse contextualized word representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2459–2471, Seattle, United States. Association for Computational Linguistics.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explain the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022a. 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 (Volume 1: Long Papers)*, pages 4331–4352.

Niccolò Campolungo, Tommaso Pasini, Denis Emelin, and Roberto Navigli. 2022b. Reducing disambiguation biases in nmt by leveraging explicit word sense information. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4824–4838.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not greek to mbert: Inducing word-level translations from multilingual bert. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56.

Bradley Hauer and Grzegorz Kondrak. 2023. Taxonomy of problems in lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9833–9844.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021. Semi-supervised and unsupervised sense annotation via translations. *arXiv preprint arXiv:2106.06462*.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *International conference on text, speech and dialogue*, pages 103–111. Springer.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. 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.

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.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic

network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural language engineering*, 5(2):113–133.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "OneSeC" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709, Florence, Italy. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Multilingual word sense disambiguation with unified sense representation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4193–4202.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Junwei Zhang, Ruifang He, Fengyu Guo, Jinsong Ma, and Mengnan Xiao. 2022. Disentangled representation for long-tail senses of word sense disambiguation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2569–2579.

A Additional Experimental Details

We present the full set of C-WLT prompts for all 18 evaluation languages from Section 4 in Table 5. We note that for the templates with a [target word], the context prior to [target word] is fed into the PLM as the prompt, and candidates in a target language are concatenated with the part after [target word] to calculate the final score of each potential translation.

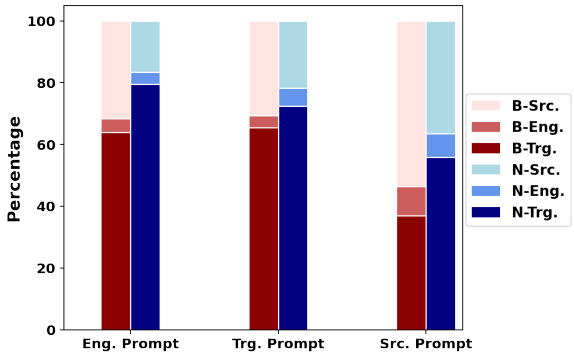

Figure 6: Proportion of top-1 predictions in different languages by prompt language. *Trg.* language predictions are the desired language choice, while *Src.* is predictions in the prompt language.

B Additional Analysis of WSD via C-WLT

Figure 6 presents the top-1 predicted languages analysis from Section 4.2.

B.1 Effect of Sense Frequency on Performance

Supervised WSD classifiers often learn to predict more commonly seen senses in the training data, which leads to stronger performance on examples of the most common sense (MCS) of words than the less common senses (LCS) (Maru et al., 2022). We test whether this behavior holds with the unsupervised WSD via C-WLT approach by evaluating performance on examples where the gold sense is the MCS of the word and those annotated with an LCS separately (Table 4).

Language	Recall				Jaccard Index			
	Bloom-7.1B		GPT-NeoX		Bloom-7.1B		GPT-NeoX	
	MCS	LCS	MCS	LCS	MCS	LCS	MCS	LCS
Basque	79.22	42.25	71.84	36.24	72.84	28.48	65.70	23.41
Bulgarian	83.54	56.38	86.79	60.13	79.04	42.98	81.18	45.96
Catalan	71.89	50.11	73.13	48.66	60.37	32.93	60.45	30.91
Chinese	75.82	49.74	76.81	52.41	66.56	35.34	66.45	36.32
Croatian	87.63	50.62	89.01	54.35	84.32	38.50	85.26	41.19
Danish	87.89	58.23	90.01	66.53	83.93	45.48	85.39	51.80
English	91.20	61.52	90.51	61.01	84.63	46.32	83.90	45.43
Estonian	79.33	44.10	82.53	51.03	75.03	33.42	77.73	36.70
French	93.07	59.81	88.35	61.14	86.25	45.92	81.84	43.90
Galician	85.83	66.13	86.54	64.39	78.37	47.00	78.56	46.21
German	89.08	60.87	87.97	63.48	84.92	43.71	84.36	47.04
Hungarian	81.31	42.24	84.73	49.50	77.06	31.30	79.72	36.84
Italian	86.62	65.78	85.68	66.54	73.84	45.81	73.85	46.28
Japanese	82.83	54.94	84.53	58.42	76.21	38.10	77.29	39.48
Korean	81.98	42.93	81.47	41.96	79.31	32.55	78.71	32.16
Slovenian	69.02	40.07	77.90	43.85	59.50	28.68	68.69	29.73
Spanish	87.81	71.12	87.12	67.94	71.83	49.75	71.72	45.25
Avg.	83.16	53.93	83.82	55.74	75.11	39.19	76.52	39.92

Table 4: Recall and Jaccard index performance of the best-ensembled WSD via C-WLT setting for the most common senses (MCS) and less common senses (LCS) of words in each evaluation language.

The results show that the gap between MCS and LCS performance is relatively large for both metrics: we observe an average difference of 28.7 and 36.3 between MCS and LCS examples for recall and Jaccard index, respectively. We also find that the size of this performance gap is consistent between the GPT-NeoX and Bloom-7.1B models. We hypothesize that this performance gap stems from unbalanced latent sense supervision in the pretraining data that is due to the natural Zipfian distribution of senses in language (Kilgarriff, 2004). This finding then highlights that even zero-shot methods extrapolating from the pretraining signals are still vulnerable to unbalanced data.

C Responsible NLP Miscellanea

This section details information from the Responsible NLP Checklist not covered elsewhere in the paper.

Intended Usage of Artifacts To the best of our knowledge, our experiments all fall within the intended use cases of the GPT-Neo and BLOOM models. We also use all data resources – the XL-WSD dataset, BabelNet, and WordNet – as originally intended (i.e., for WSD modeling and evaluation).

D Full Experimental Results

We provide the per-langauge results for the EN-ZH (Figure 7), EN-FR (Figure 8), and EN-ES (Figure 9) contextual WLT experiments. In these figures, the top row relays results of the zero-shot contextual WLT accuracies on GPT and BLOOM family models of different sizes. The bottom left figure indicates the average NLL of all correct and incorrect words across models in the contextual WLT analysis, with labels of the NLL ratio of incorrect to correct translations; the bottom right plots the impact of adding context to WLT on translation (trans.) and disambiguation (disam.) errors.

Additionally, Table 6 reports the Bloom-3B and Bloom-7.1B results for the target language ablation and ensembling experiments from Section 4.2.

Lang.	Prompt Template
English	In the sentence “<sentence>”, the word “<source word>” is translated into <target langage> as
Spanish	En la oración “<sentence>”, la palabra “<source word>” se traduce al <target lang> como
Chinese	在“<sentence>”这句话中, “<source word>”这个词翻译成<target language>为
Catalan	A la frase “<sentence>”, la paraula “<source word>” es tradueix <target lang> com a
Basque	“<sentence>” esaldian, “<source word>” <target lang> [target word] gisa itzultzen da
German	In dem Satz „<sentence>“ bedeutet das Wort „<source word>“ ins <target lang> als
Estonian	Lauses “<sentence>” tõlgitakse sõna “<source word>” <target lang> keelde kui
French	Dans la phrase “<sentence>”, le mot “<source word>” se traduit en <target lang> par
Bulgarian	В изречението „<sentence>“ думата „<source word>“ се превежда на <target lang> като
Croatian	U rečenici “<sentence>”, riječ “<source word>” prevedena je na <target lang> kao
Danish	I sætningen “<sentence>” oversættes ordet “<source word>” til <target lang> som “
Dutch	In de zin “<sentence>” vertaalt het woord “<source word>” zich in het <target lang> als “
Galician	Na frase “<sentence>”, a palabra “<source word>” tradúcese ao <target lang> como
Hungarian	A “<sentence>” mondatban fordítsa le a “<source word>” szót <target lang>
Italian	Nella frase “<sentence>”, la parola “<source word>” si traduce in <target lang> come
Japanese	「<sentence>」という文で、「<source word>」という単語は<target lang>に訳すと [target word] となります
Slovenian	V stavku “<sentence>” se beseda “<source word>” v <target lang> prevede kot
Korean	“<sentence>”이라는 문장에서 “<source word>”이라는 단어는 <target lang> [target word]로 번역됩니다

Table 5: C-WLT templates we used in the experiment for different prompt languages.

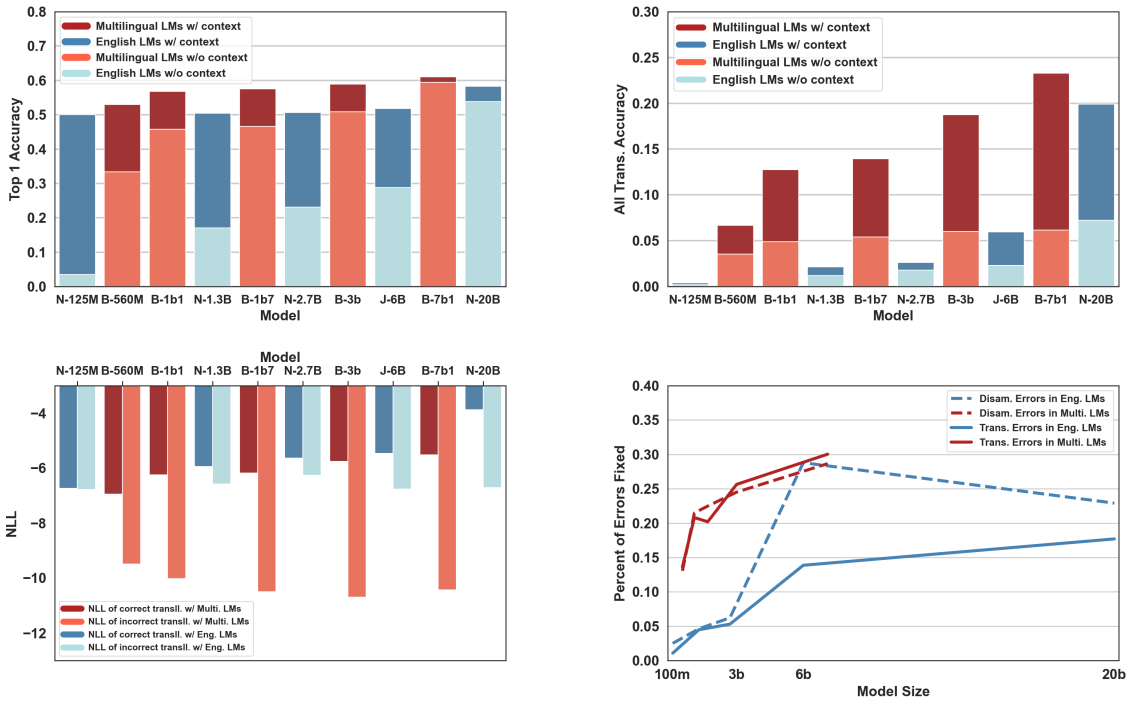

Figure 7: C-WLT results for Chinese. N: GPT-Neo, B: BLOOM, J: GPT-J

Target Lang.	Recall		Jaccard Index		Delta	
	B-3B	B-7.1B	B-3B	B-7.1B	B-3B	B-7.1B
English	63.60	63.62	51.83	52.32	10.1	9.7
Spanish	69.58	69.86	52.28	52.31	15.7	15.6
Chinese	68.77	69.96	57.43	58.27	4.1	4.1
Russian	65.06	65.68	53.75	54.39	9.4	9.4
Finnish	55.01	56.52	47.73	48.73	6.9	6.5
Best Setting*	68.62	69.45	57.42	58.24	8.7	8.2
All 5 Joint	63.95	65.03	55.42	56.35	6.5	6.4

Table 6: The average zero-shot recalls and Jaccard Index (%) of all 18 source languages in the XL-WSD dataset for the different target language settings for the BLOOM family PLMs. *The best setting is the joint English, Chinese, and Russian.

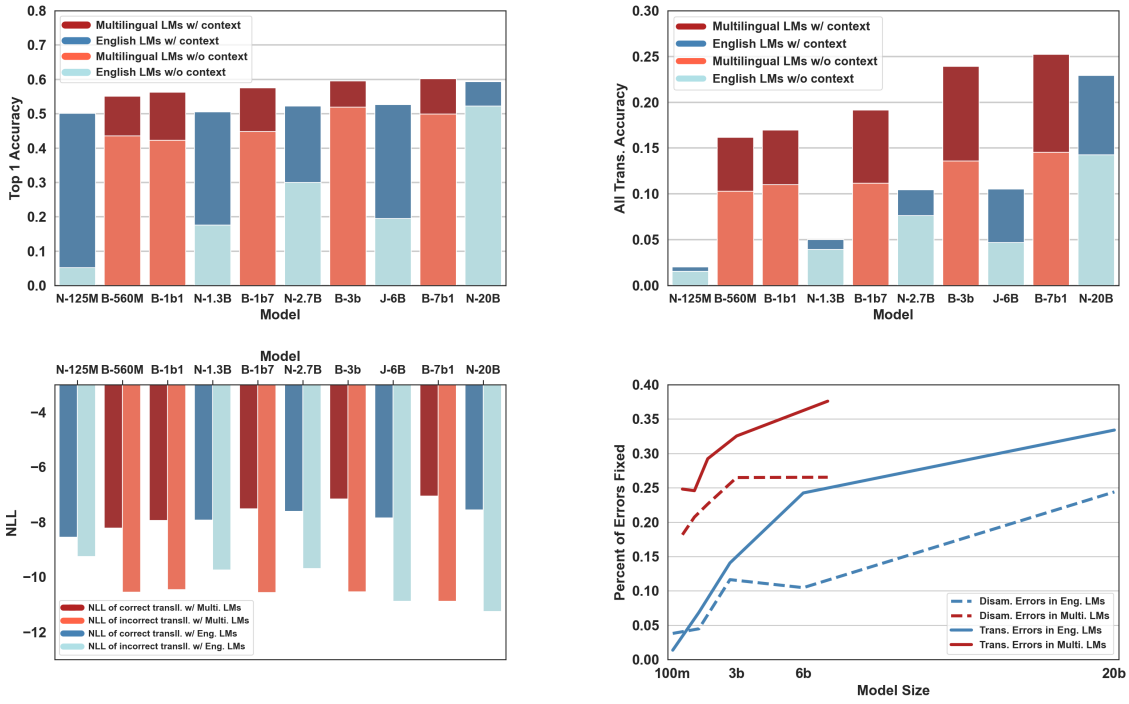

Figure 8: C-WLT results for French. N: GPT-Neo, B: BLOOM, J: GPT-J

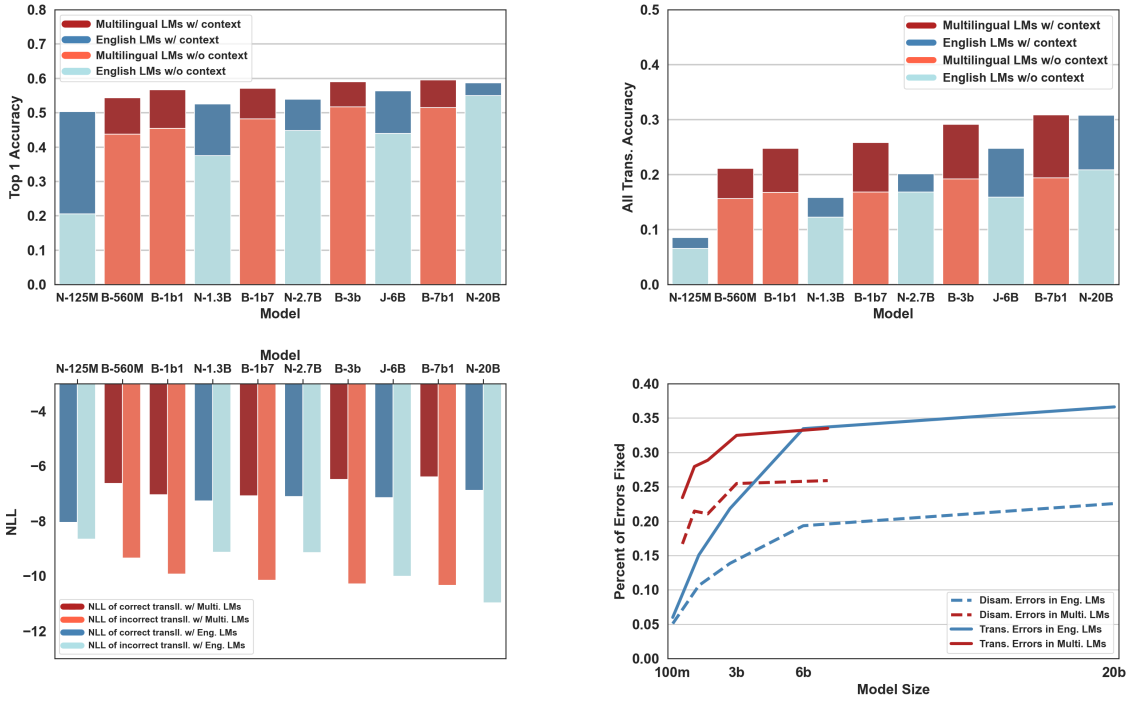

Figure 9: C-WLT results for Spanish. N: GPT-Neo, B: BLOOM, J: GPT-J