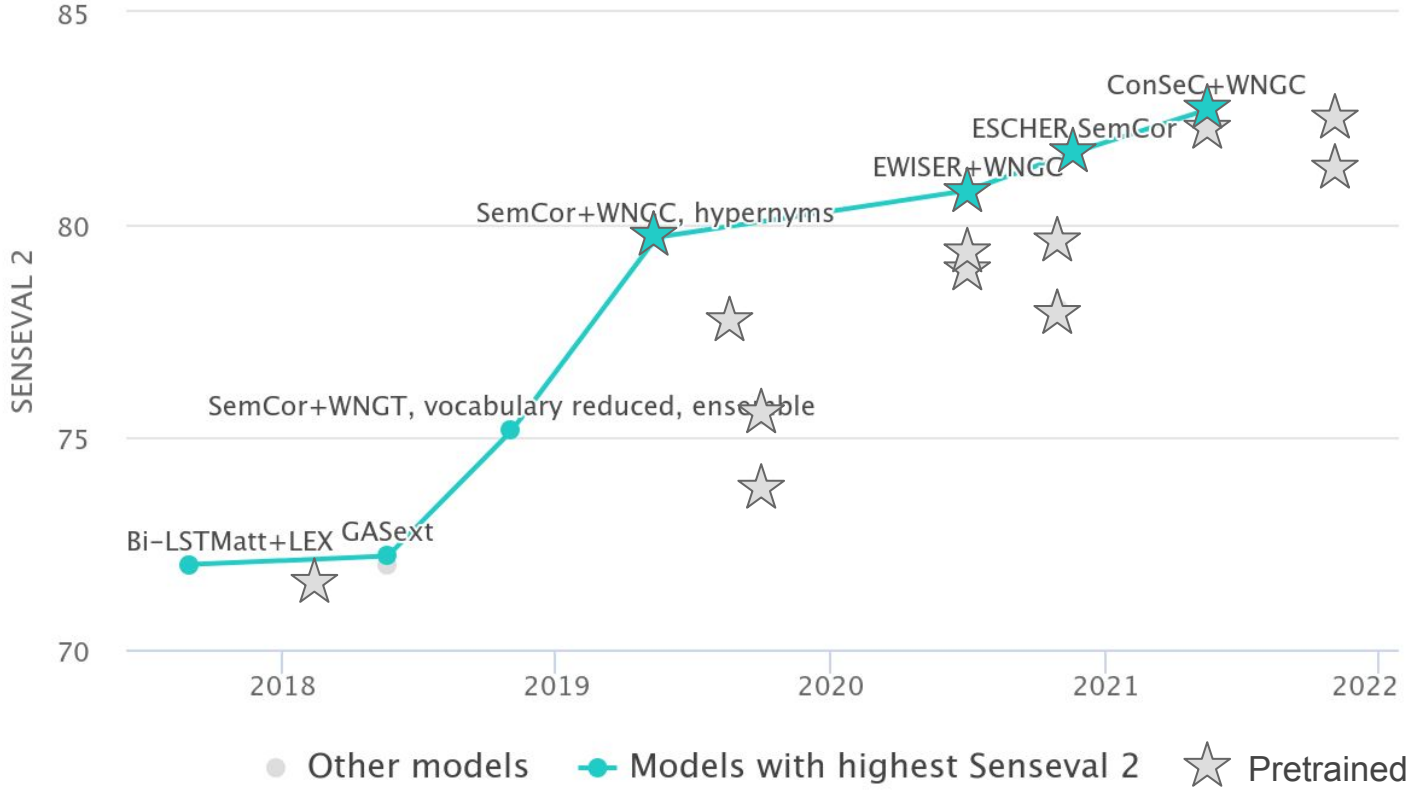
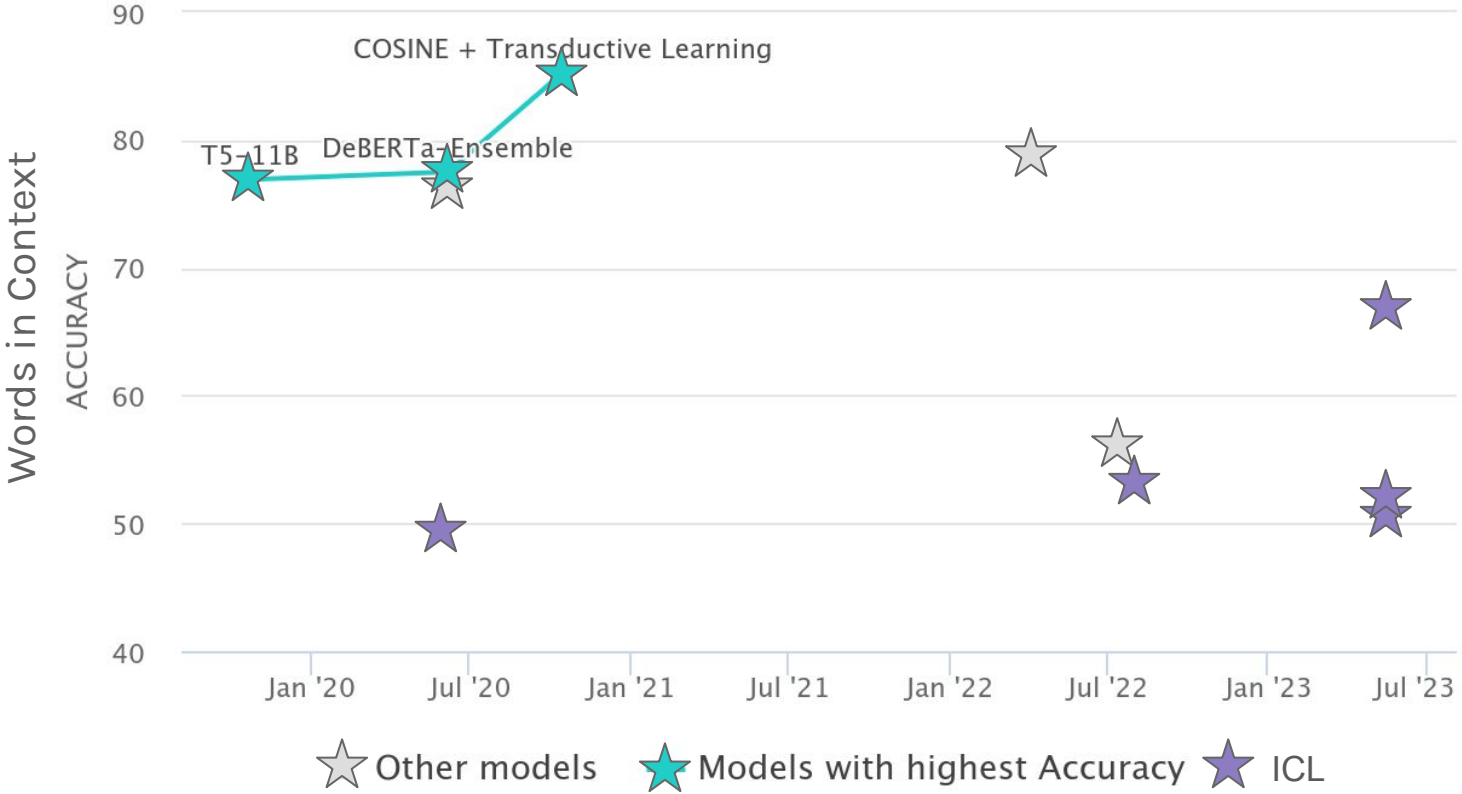
Translate to Disambiguate:
Zero-shot Multilingual Word Sense Disambiguation with Pretrained Language Models

Haoqiang Kang*, Terra Blevins*, and Luke Zettlemoyer

Pretrained Language Models are the backbone of SOTA word sense disambiguation models



Pretrained Language Models don't recover word sense with in-context learning



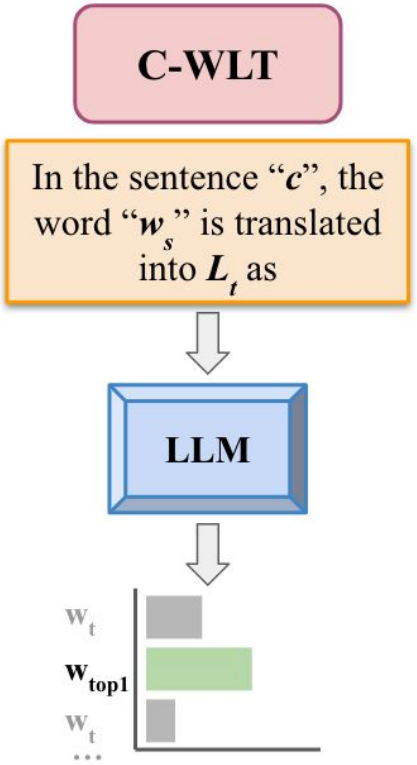
An outfitter provided everything needed for the safari.
Before his first walking holiday, he went to a specialist outfitter to buy some boots.
question: Is the word 'outfitter' used in the same way in the two sentences above?
answer:

GPT-3 WiC Prompt

Brown et al., NeurIPS 2020

```
An outfitter provided everything needed for the safari.
Before his first walking holiday, he went to a specialist outfitter to buy some boots.
question: Is the word 'outfitter' used in the same way in the two sentences above?
answer:
```

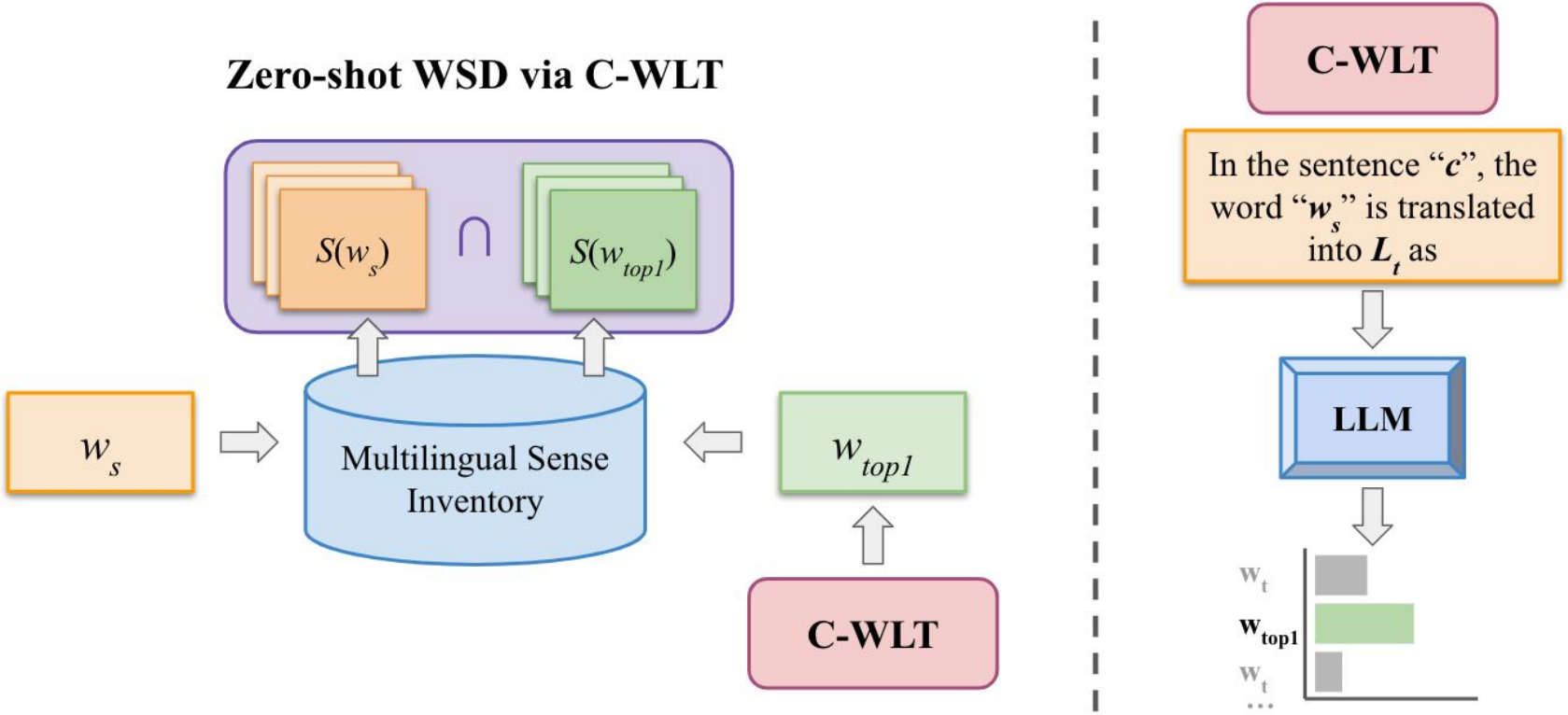
GPT-3 WiC Prompt



Contextual Word-Level Translation (C-WLT)

a more natural method to prompt LMs for word sense knowledge

Extend Contextual Word-Level Translation (C-WLT) into zero-shot approach for multilingual WSD

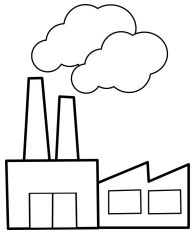


Word-level Translation (WLT)

Source Word Target Language

WLT: The word “plant” is translated into Chinese as ___.

a) 

b) 

c) 

Contextual Word-level Translation (WLT)

Source Word Target Language

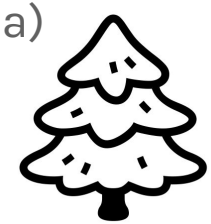
WLT: The word “plant” is translated into Chinese as ___.

Context

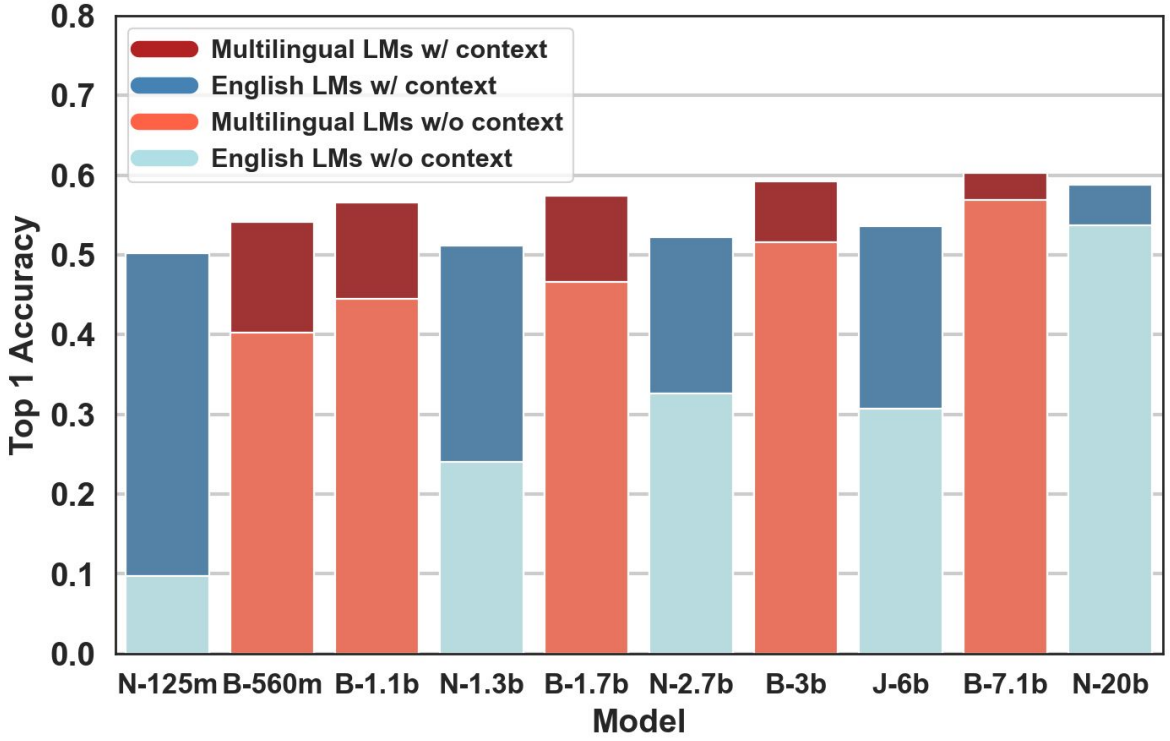
Contextual WLT: In the sentence “The plant sprouted a new leaf”, the word “plant” is translated into Chinese as ___.

Source Word Target Language

a)

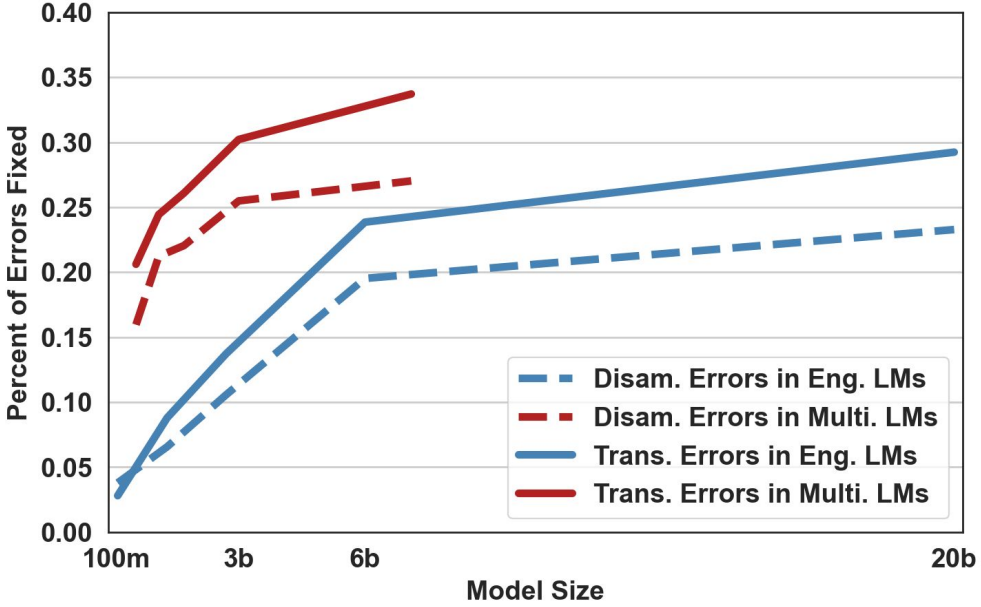


Results: Adding Context Improves WLT Accuracy



Results

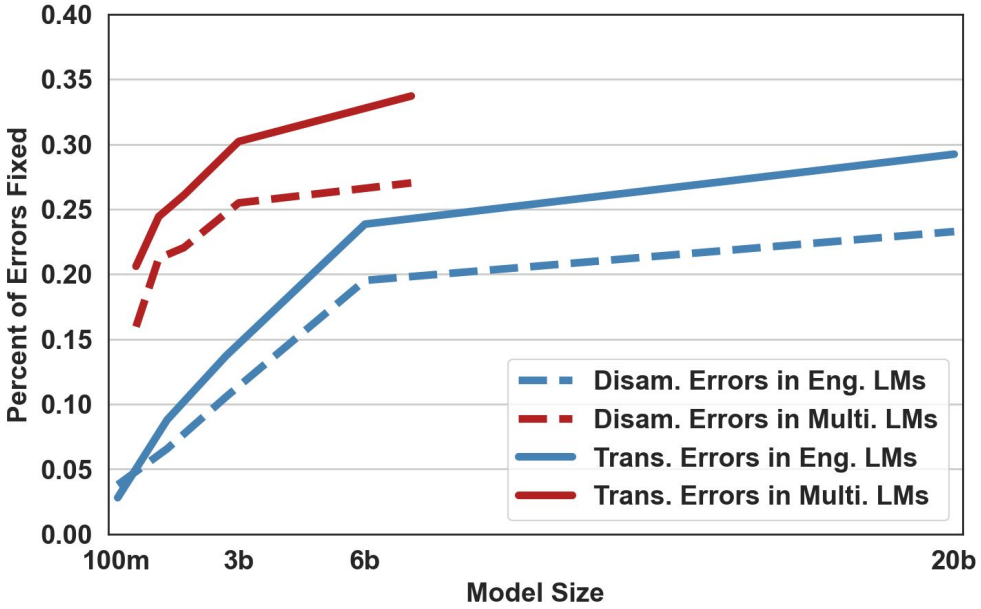
- **Larger** models benefit more than smaller ones from using contextual information.



Results

- **Larger** models benefit more than smaller ones from using contextual information.

- Context helps correct **translation errors** as well as disambiguate the sense



What does this tell us?

Contextual word-level translation (**C-WLT**) can differentiate different meanings of a word in the source language

What does this tell us?

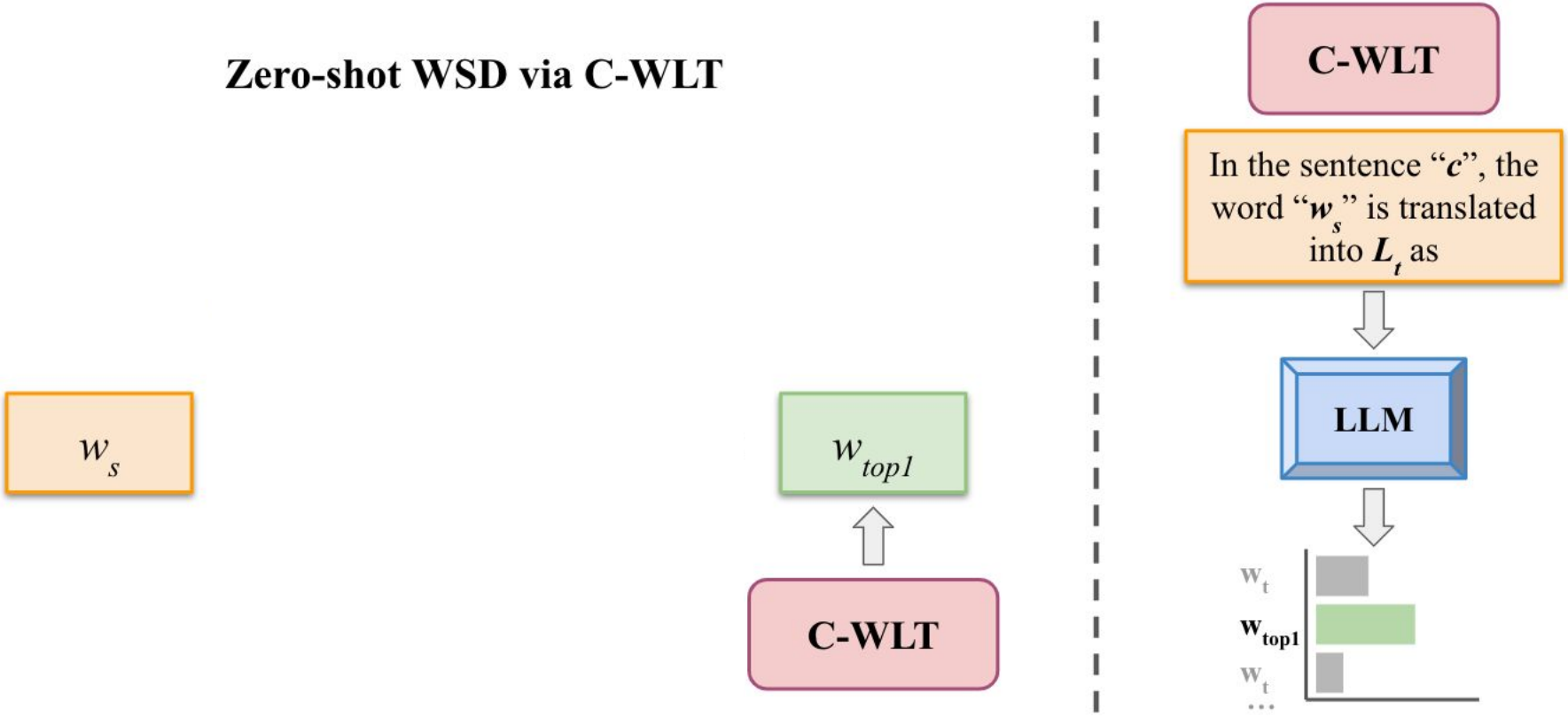
Contextual word-level translation (C-WLT) can differentiate different meanings of a word in the source language

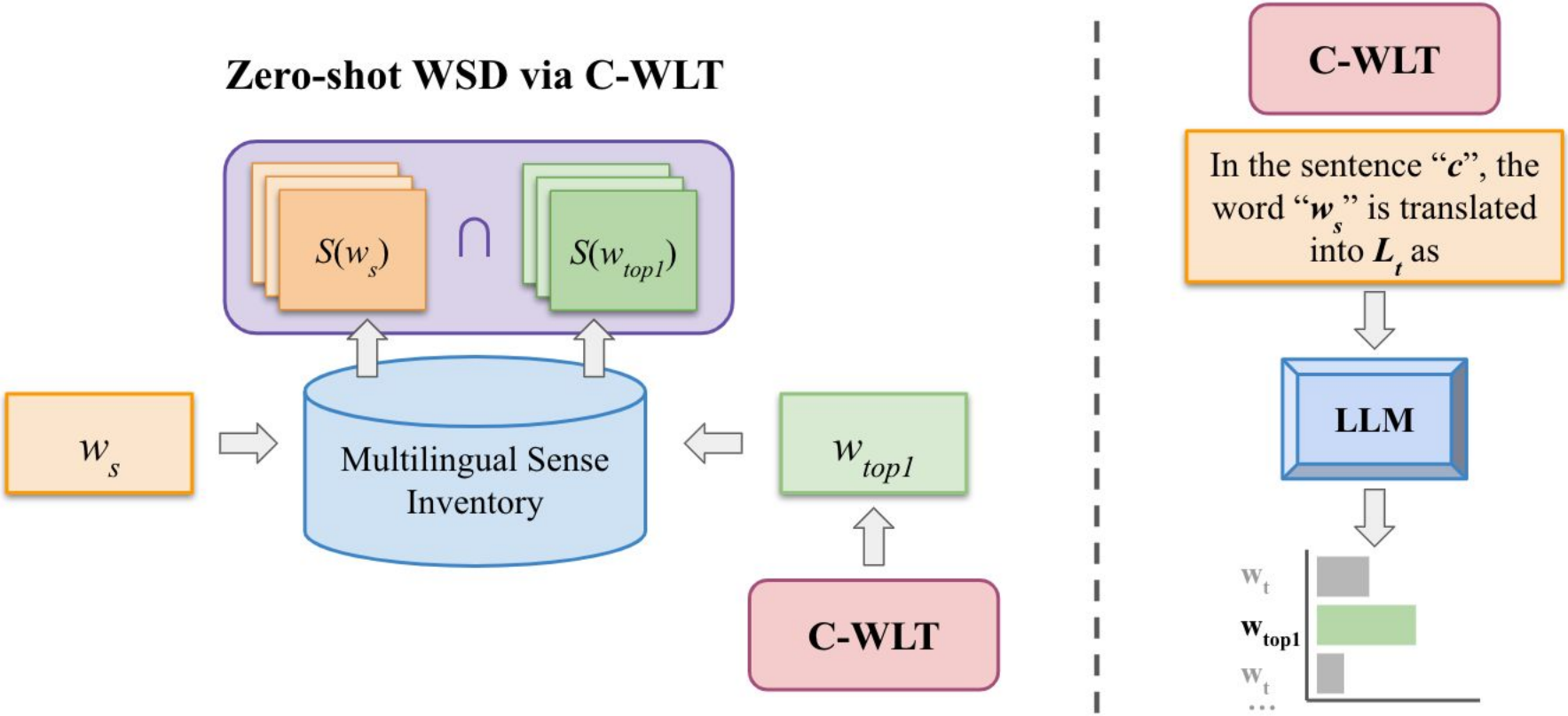
Can we use C-WLT to perform Word Sense Disambiguation?

Zero-shot WSD via C-WLT

w_s

Zero-shot WSD via C-WLT




Zero-shot WSD via C-WLT
$S(w_s)$
$S(w_{top1})$
w_s
Multilingual Sense Inventory
w_{top1}
C-WLT
C-WLT
In the sentence "c", the word "w_s" is translated into L_t as
LLM
w_t
w_{top1}
w_t
...

WSD Experimental Setup

- Evaluation on **XL-WSD** dataset (18 source languages)
- 5 target languages (English, Chinese, Russian, Spanish, and Finnish)

WSD Experimental Setup

- Evaluation on **XL-WSD** dataset (18 source languages)
- 5 target languages (English, Chinese, Russian, Spanish, and Finnish)

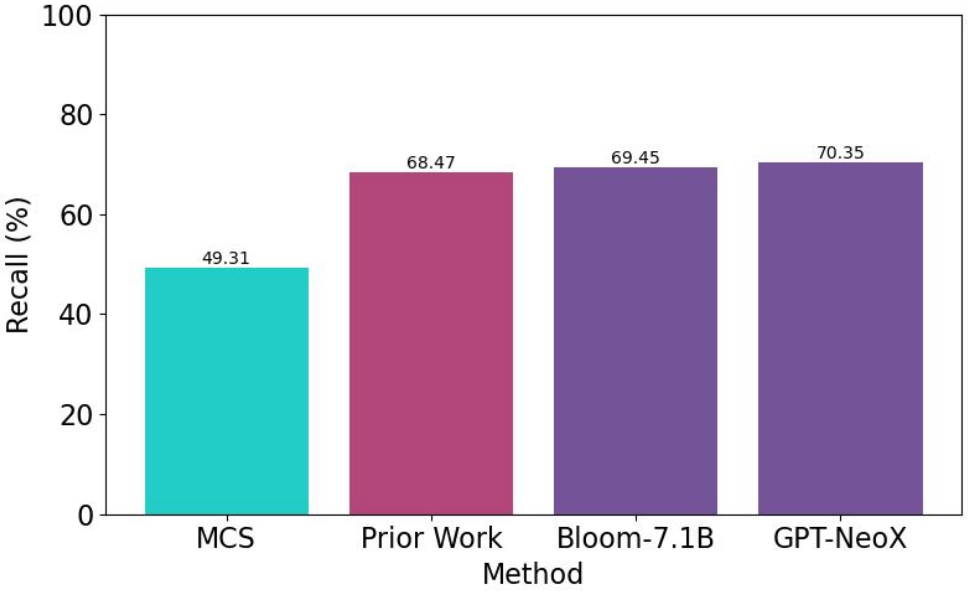
Target Language Ensembling

Given a source word w_s and a set of target languages T:

1. $S(T) = \{S(w^t_{top1}): t \in T\}$
2. $S(T)' = \{max(s) \in S(T)\}$
3. Predicted sense set = $S(T)' \cap S(w_s)$

WSD Results

WSD via C-WLT achieves higher recall than the **prior supervised works** in 11 out of the 18 source languages.



Why Recall?

Recall

How often the predicted label set contains ***at least one*** *of the gold annotations for a given example.*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(L_i^{pred} \cap L_i^{gold} \neq \emptyset)$$

Jaccard Index

Percentage of overlap between the true and predicted label sets.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|L_i^{pred} \cap L_i^{gold}|}{|L_i^{pred} \cup L_i^{gold}|}$$

WSD Results

- There is a **tradeoff** between the Jaccard index and recall...

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

WSD Results

- There is a **tradeoff** between the Jaccard index and recall...

- Mitigated by using **dissimilar** target languages to the source and ensembling diverse languages.

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

Effect of Fine-grained Senses

- WSD via C-WLT predicts **sets** of senses…
- …but most words in XL-WSD are annotated with one sense
- Missing senses are a major source of annotation error in WSD datasets (Maru et al., ACL 2022)

Effect of Fine-grained Senses

Sentence: 广播还没说完，各班的同学早已冲出教室 (*Before the broadcast was finished, students from all classes had already rushed out of the classroom.*)

XL-WSD Sense: "*Be broadcast*".
Missing Sense: "*Broadcast over the airwaves, as in radio or television.*"

Effect of Fine-grained Senses

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**

Sentence: 广播还没说完，各班的同学早已冲出教室 (*Before the broadcast was finished, students from all classes had already rushed out of the classroom.*)

XL-WSD Sense: "*Be broadcast*".
Missing Sense: "*Broadcast over the airwaves, as in radio or television.*"

Manual reannotation of Chinese WSD data verifies that many words (~44%) have missing, related senses

Takeaways

- **C-WLT** recovers word sense from LMs through prompting
- **WSD via C-WLT** uses translation to perform zero-shot WSD with LMs

Takeaways

- **C-WLT** recovers word sense from LMs through prompting
- **WSD via C-WLT** uses translation to perform zero-shot WSD with LMs
- These methods **demonstrate LM's knowledge of underlying word senses** and **present new approaches for WSD in low-resource settings**

Takeaways

Check out the paper!

- **C-WLT** recovers word sense from LMs through prompting
- **WSD via C-WLT** uses translation to perform zero-shot WSD with LMs
- These methods **demonstrate LM's knowledge of underlying word senses** and **present new approaches for WSD in low-resource settings**

Questions?

http://mk322.github.io

haoqik88@gmail.com

http://blvns.github.io

blvns@cs.washington.edu