Code-Switching with Word Senses for Pretraining in Neural Machine Translation

CALCS 2023

Vivek Iyer[1], Edoardo Barba[2], Alexandra Birch[1], Jeff Pan[1], Roberto Navigli[2]

[1]The University of Edinburgh [2]Sapienza University of Rome

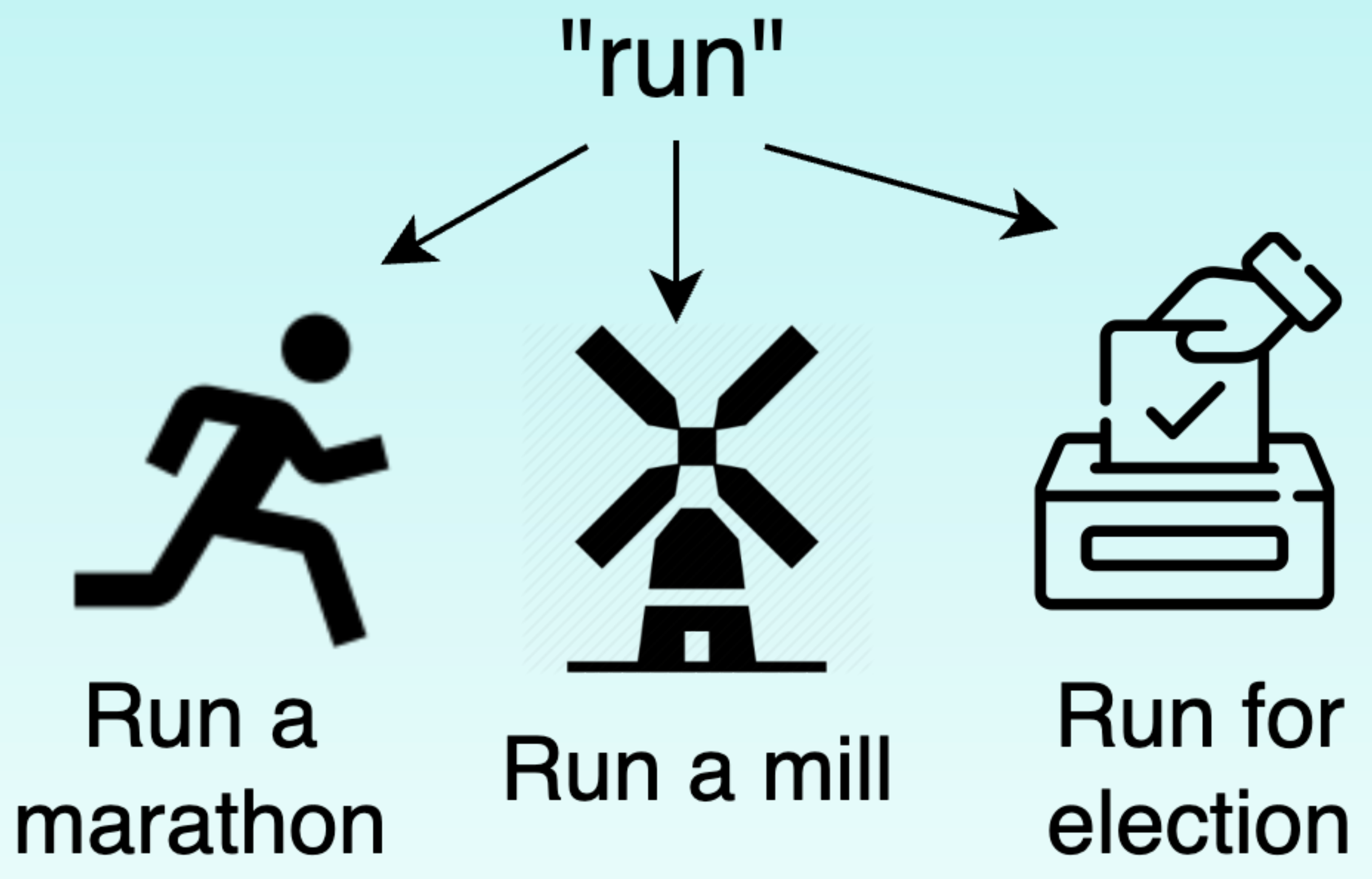
vivek.iyer@ed.ac.uk

Agenda

- Motivate the problem
  - Lexical ambiguity in NMT
- Problems with current NMT pretraining paradigm
- Discuss “code-switched pretraining”
- Distinguish from human code-switching
- Explain our approach: code-switching with word senses
- Discuss (qualitative + quantitative) results
- Finally, mention some applications

The Problem

- Lexical Ambiguity is a fundamental challenge in MT
    - “Problem of multiple meanings” (Weaver, 1947)



Motivation

- Many modern-day NMT systems struggle with WSD, and display several biases against rare or polysemous word senses (Campolungo et al., 2022)

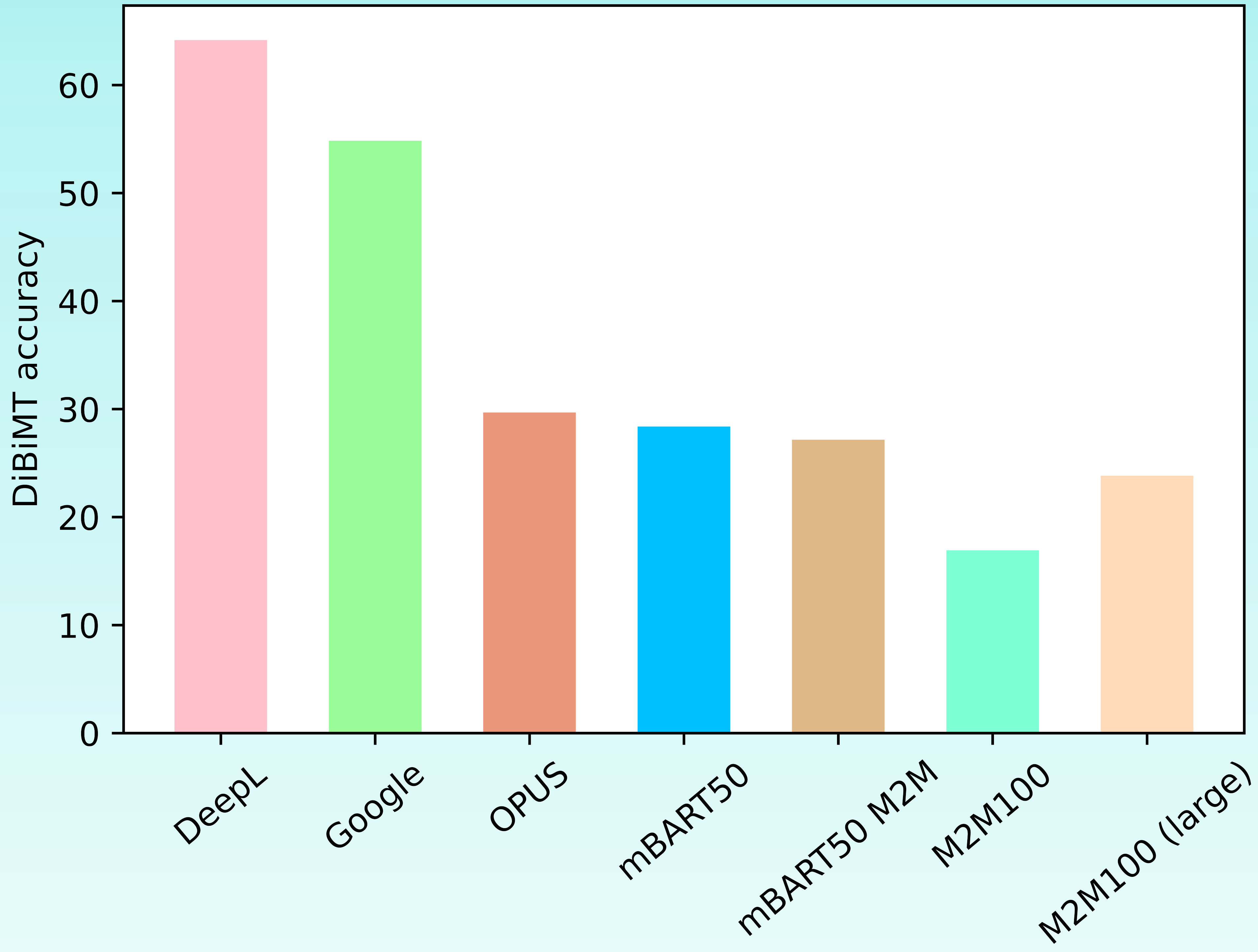


Figure 1: Disambiguation accuracy of some well-known MT systems [3]

Why?

- We hypothesise the answer lies in “sense-agnostic” NMT pretraining! Particularly, **code-switched pretraining**

Code-Switched Pretraining: A review

- Along with masked denoising (eg. mBART), one of the most common pretraining techniques in NMT over the last 4 years [1][2][3][4][5][6][7]
- Synthetic Code-Switching of words in a sentence with lexical translations. Random & Multilingual
- Aligned Augmentation (AA) [3]: Noteworthy work in this area
- NMT models are pretrained to “de-codeswitch” these sentences.
- Resulting models show strong cross-lingual convergence; huge improvements in MT scores

1	Original (En)	One more point is lost in this debate: that the EU is proposing far fewer rules now.
	AA	One высокого πόντος той perduti العام tento diskusijos : tuo cette EU is soovitab 遠く 低い регламент घंटे .
2	Original (En)	" If we don 't win , there will be some inquiries of why we haven't , " Graves told BBC Radio Leeds.
	AA	" If noi annetada 't ויטוריה , そこ хочу jet sometime αιτήσεις seine kuna bize haven't , " Graves erzählte BBC Radio Leeds.

Source: Figure 6, Pan et al., 2021. Contrastive Learning for Many-to-many Multilingual Neural Machine Translation.

So, what's the problem?

- **Polysemy!** => Lexical translations randomly chosen
- "**Sense-agnostic pretraining**": Synthetic code-switching happens at the word-level, not the sense-level
- Potential cause for WSD biases/failures?
- We propose "**Sense-pivoted pretraining**" => Move code-switching to the sense level, rather than the word level

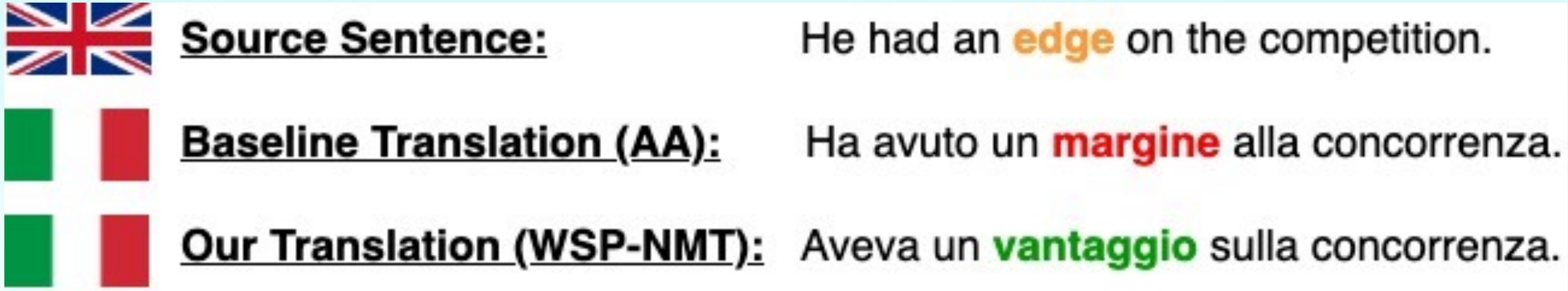


Figure 3: AA vs WSP-NMT. ***Margine=edge, vantaggio=advantage***

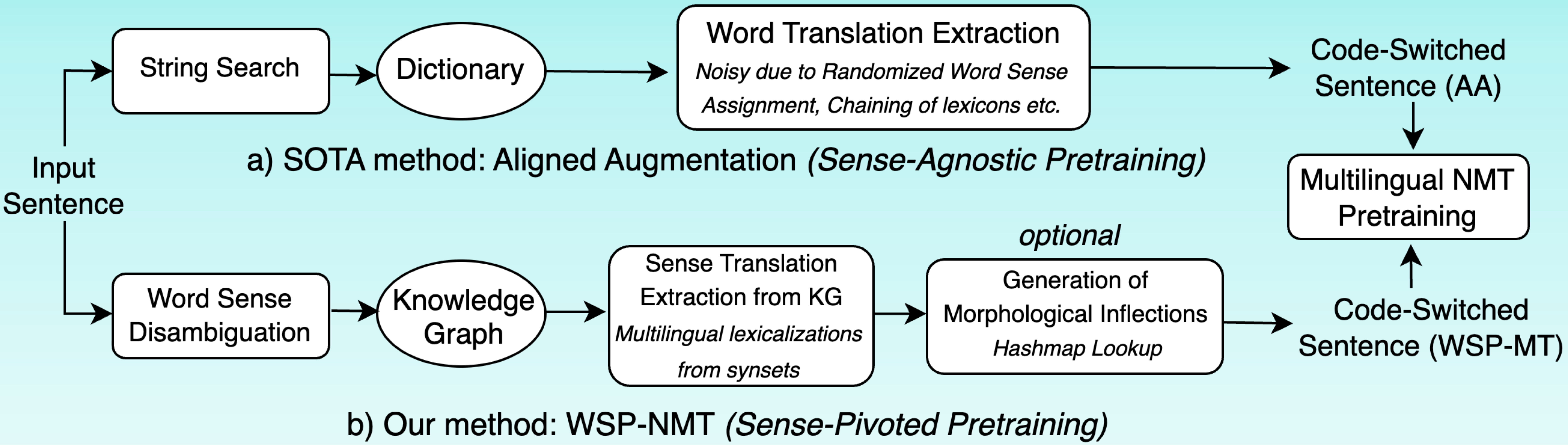
A note on code-switching

- **What does this presentation discuss?**
  - Technique for generating synthetic code-switched data
- **Why are we generating this data?**
  - For pretraining general-purpose multilingual NMT models
  - We **do not** seek to evaluate on code-switched MT
- **How would this differ from human code-switching?**
  - Does not follow definitive rules/patterns. Quite random, massively multilingual
  - Purpose is to teach NMT systems lexical translation!

Contributions

- We propose Word Sense Pretraining for Neural Machine Translation (WSP-NMT), using WSD + KG for code-switching
    - WSD-based code-switching > lexicon-based code-switching
    - KG in NMT pretraining => less errors, better quality
- Experiments in data and resource-constrained scenarios
- Evaluate disambiguation performance on DiBiMT MT benchmark

Approach



In NMT pretraining, CS sentence is aligned with original sentence w/ contrastive loss (+ cross entropy)

Experimental Setting

- Primary baseline: Aligned Augmentation (AA) [3]

- Multilingual NMT pretraining on Romance languages (En-Es, En-Fr, En-It, En-Ro).
    - Parallel + mono data
    - En-Pt is zero-shot.
    - CS done with AA and WSP-NMT; shuffled

- WSD systems:
    - AMuSE-WSD (cheap, yet competitive)
    - ESCHER (slow, but prev. SOTA on English WSD)

Main Results

☑ Consistent gains over AA

☑ Better WSD (ESCHER) = better MT quality. But AMuSE-WSD is effective too! (2.3x cheaper)

☑ Morph. Inflection prediction for word senses helps! {gender, tense} agreement

☑ Lower-resourced En-Ro (5x less data) gains the most!!

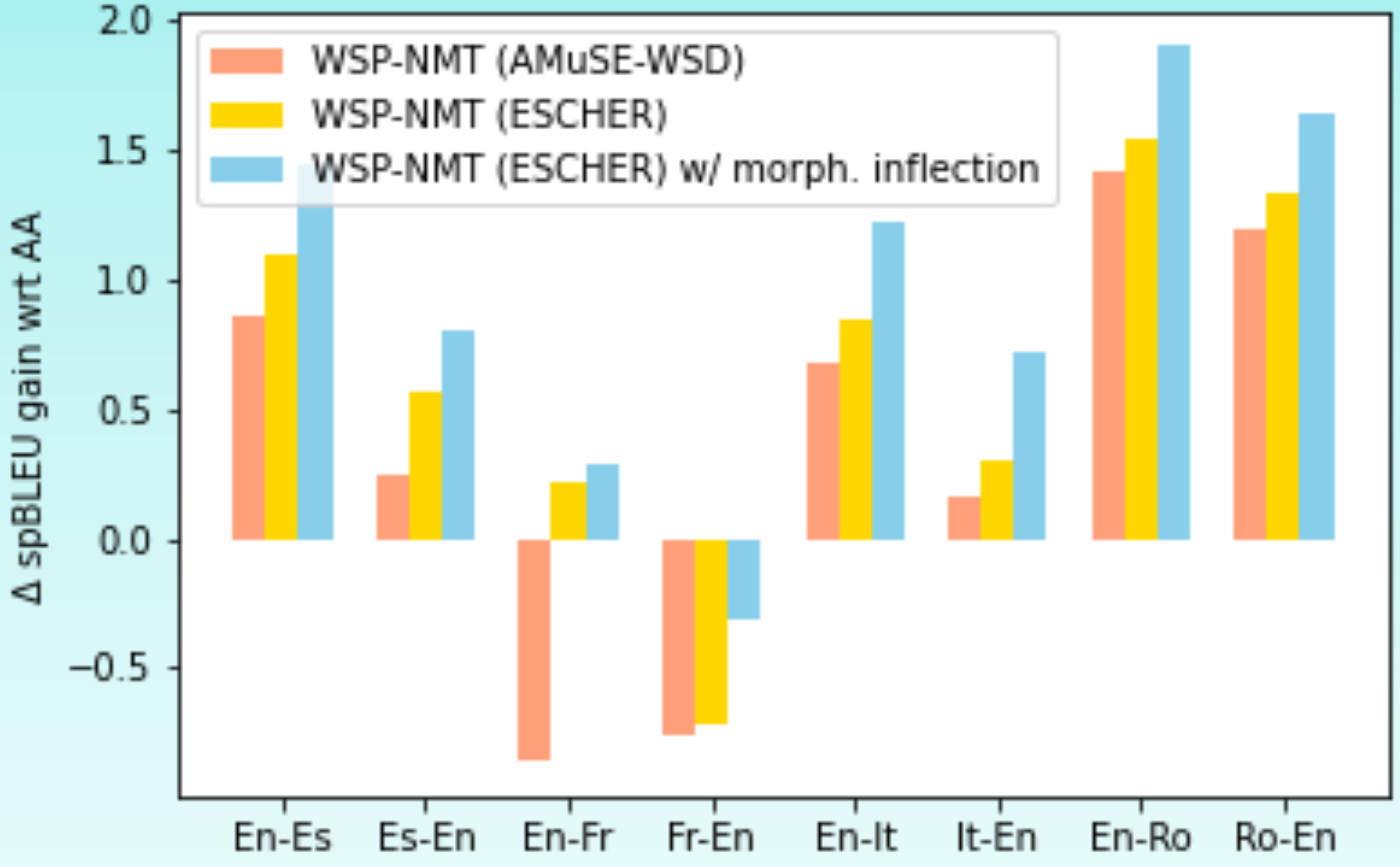
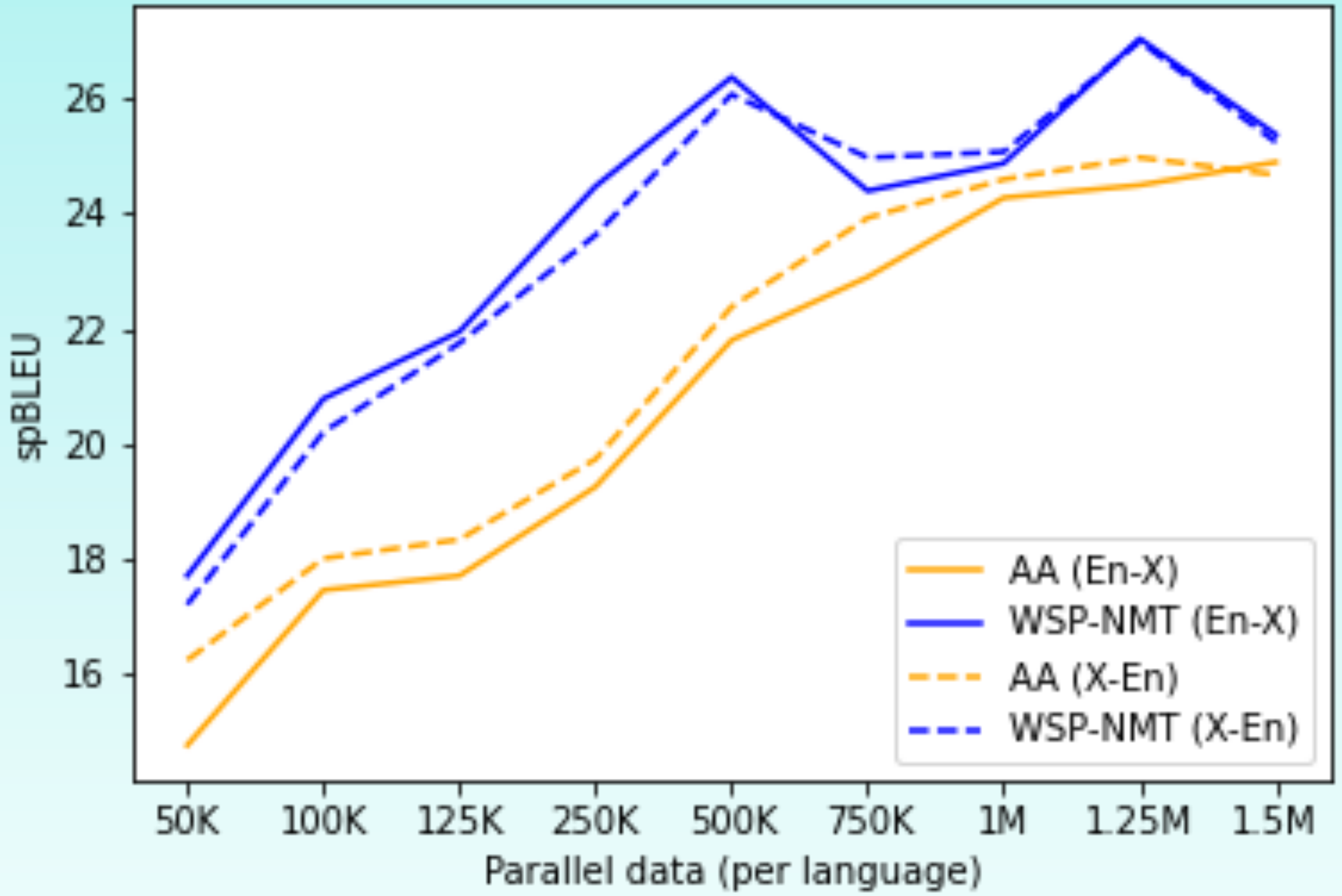


Figure 4: Overall MT quality (spBLEU) gains for WSP-NMT over AA

Resource-Constrained Settings

a) Data quantity vs performance



Highly effective in low & medium data (<750K parallel sents) settings!

b) Zero-shot MT

Table 1: Zero-shot spBLEU

Baseline	En-Pt	Pt-En
AA	2.92	6.88
WSP-NMT	**3.60**	**8.52**

Enhanced multilingual convergence = Significant zero-shot gains

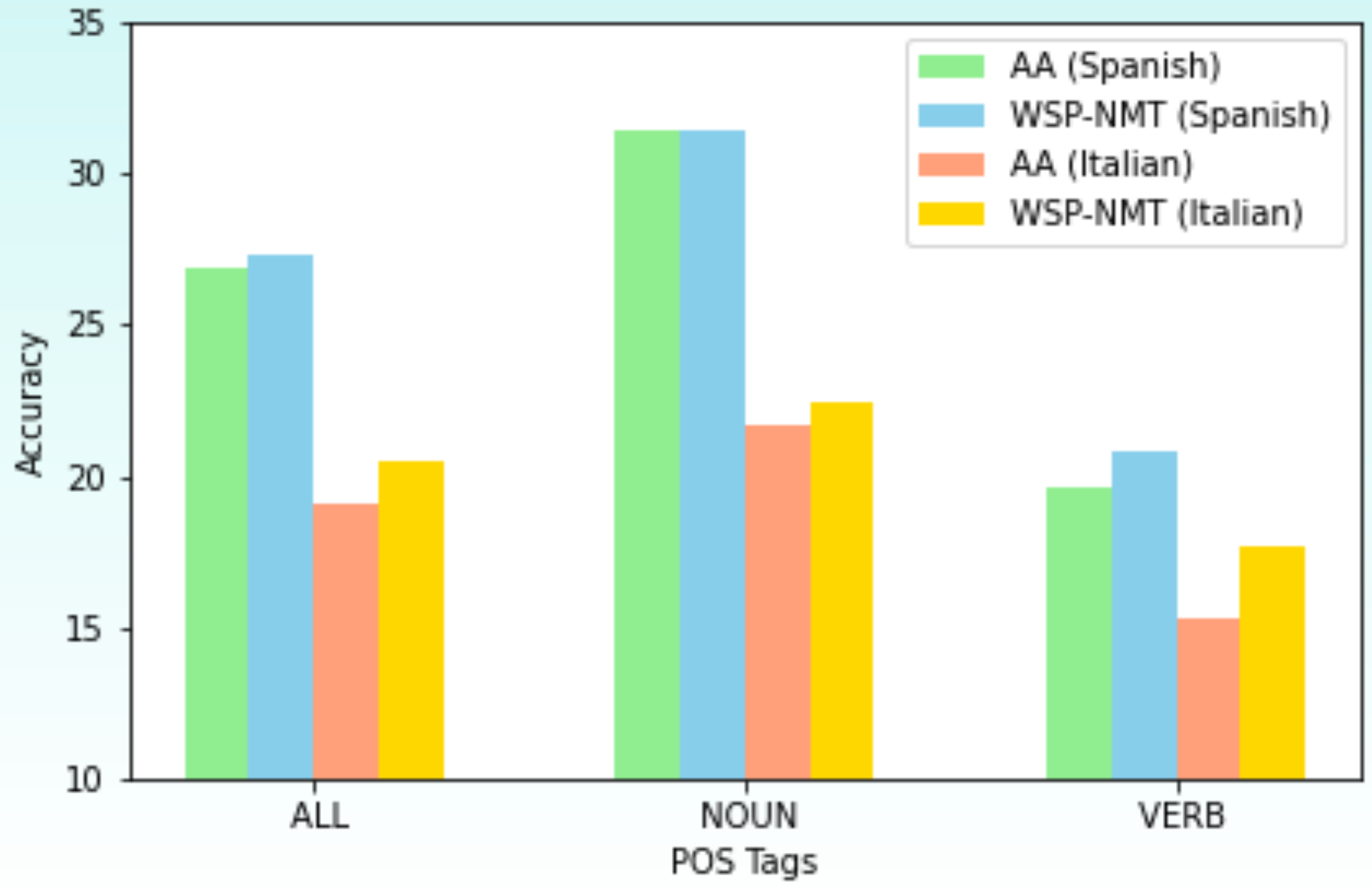
Scaling to Under-Represented Languages (Zero-shot WSD)

- Multilingual NMT for Indo-Iranian Languages (En-Hi, En-Fa)
- Zero-shot AMuSE-WSD
- No improvements observed :(
- Rooted in unavailability of disambiguation resources for training
  - Direction for future research
  - Low amount of annotated data should suffice!

Baseline	En-X	X-En
AA	22.79	20.49
WSP-NMT	22.71	20.23

Disambiguation Results

- DiBiMT ambiguity benchmark for MT
- 500 sentences, with 1 ambiguous word
- Accuracy = % Good Translations/ (% Good + % Bad) Translations
- Accuracy (ALL) ↑, Accuracy (NOUN) ≈, Accuracy (Verb) ↑ ↑



Verb Disambiguation Examples

Figure 5a. “trasformato” = “transformed”
“fatto” = “made” (i.e. made a good profit)

Source:	The company turned a good profit after a year.
AA:	L'impresa ha trasformato un buon profitto dopo un anno.
WSP-NMT:	La società ha fatto un buon profitto dopo un anno.

Figure 5b. “adeguare” = “adapt”/“adjust”
“stanziare” = “allocate” (eg. to allocate funds)

Source:	To appropriate money for the increase of the navy.
AA:	Per adeguare il denaro per l'aumento della tassa.
WSP-NMT:	Per stanziare fondi per l'aumento dell'imbarcazione.

Figure 5c. “Aveva dovuto tornare” = “had to return”
“tornato indietro” = “move (or run) back”

Source:	The player had to backpedal before catching the ball.
AA:	Il giocatore aveva dovuto tornare prima di catturare la balla.
WSP-NMT:	Il giocatore era tornato indietro prima di prendere la palla.

Conclusion

- **Advantages:**
  - More reliability with KG, better quality MT, less errors
  - Super useful in low/medium data settings!

Disadvantages:
  - Need WSD resources (Well-resourced languages)

Applications:
  - Domain-specific translation
  - Information-centric domains
  - (Potentially) better CS translation?

THANK YOU!

Questions are unambiguously welcome :)

References

[1] Yang, Z., Hu, B., Han, A., Huang, S., & Ju, Q. (2020, November). CSP: code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2624-2636).

[2] Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., & Li, L. (2020). Pre-training multilingual neural machine translation by leveraging alignment information.

[3] Pan, X., Wang, M., Wu, L., & Li, L. (2021). Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 244–258)*

[4] Li, P., Li, L., Zhang, M., Wu, M., & Liu, Q. (2022). Universal conditional masked language pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6379–6391)*

[5] Reid, M., & Artetxe, M. (2022). Paradise: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 800-810)*

[6] Jones, A., Caswell, I., Saxena, I., & Firat, O. (2023). Bilex Rx: Lexical Data Augmentation for Massively Multilingual Machine Translation.

[7] Iyer, V., Oncevay, A., & Birch, A. (2023, May). Exploring Enhanced Code-Switched Noising for Pretraining in Neural Machine Translation. In *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 954-968).