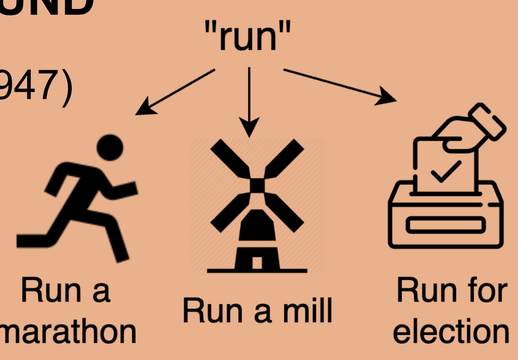
WSP-NMT: Code-Switching with Word Senses for Pretraining in Neural Machine Translation

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

1. BACKGROUND

- Lexical ambiguity in MT (Weaver, 1947)
- Modern NMT systems struggle with WSD biases
- We re-examine NMT pretraining



2. CODE-SWITCHED PRETRAINING (CSP)

- Popular NMT pretraining approach, eg. AA (Pan et al., 2021)
- Synthetic Code-Switching: Words <=> Lexical Translations
- **"Sense-agnostic"** pretraining!!

Original sentence:	"If we don 't win , there will be some inquiries of why we haven't , " Graves told BBC Radio Leeds
AA-noised sentence:	" If noi annetada 't ויטוריה , そこ хочу jet sometime αιτήσεις seine kuna bize haven't , " Graves erzählte BBC Radio Leeds.

Fig 1: Sourced from Figure 6, Pan et al., 2021

3. MOTIVATION

Idea: Disambiguate, then Code-Switch with word ***sense*** translations!

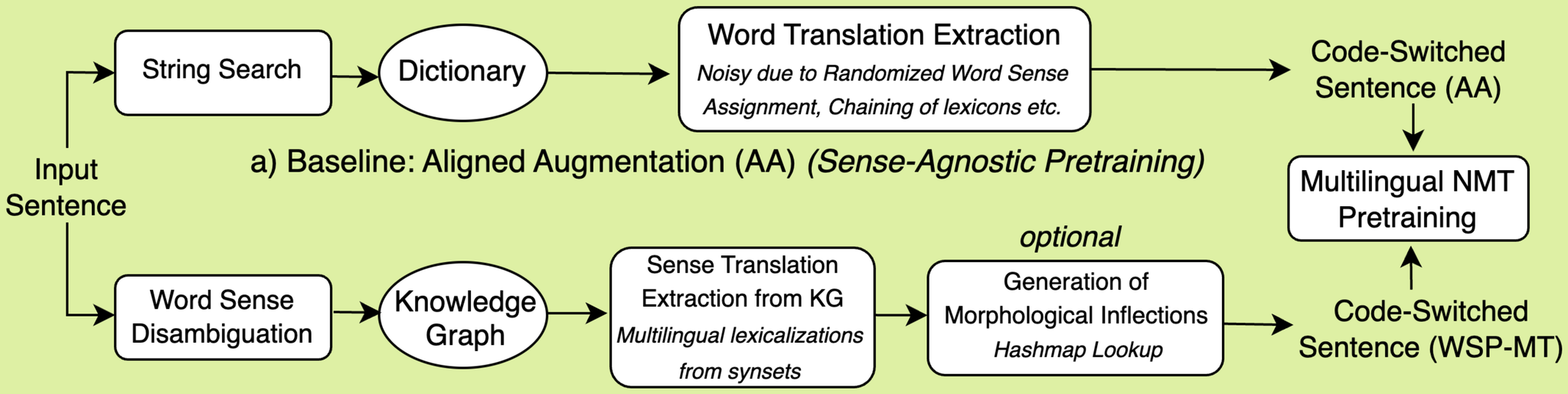
Source Sentence:	He had an **edge** on the competition.
Baseline Translation (AA):	Ha avuto un **margine** alla concorrenza.
Our Translation (WSP-NMT):	Aveva un **vantaggio** sulla concorrenza.

Fig 2: AA vs WSP-NMT. *Margine=edge, vantaggio=advantage*

4. CONTRIBUTIONS

1. **Sense-pivoted pretraining** can improve overall MT quality and WSD performance
2. KGs + mNMT pretraining = better {reliability, accuracy}
3. Super effective in data-constrained scenarios!

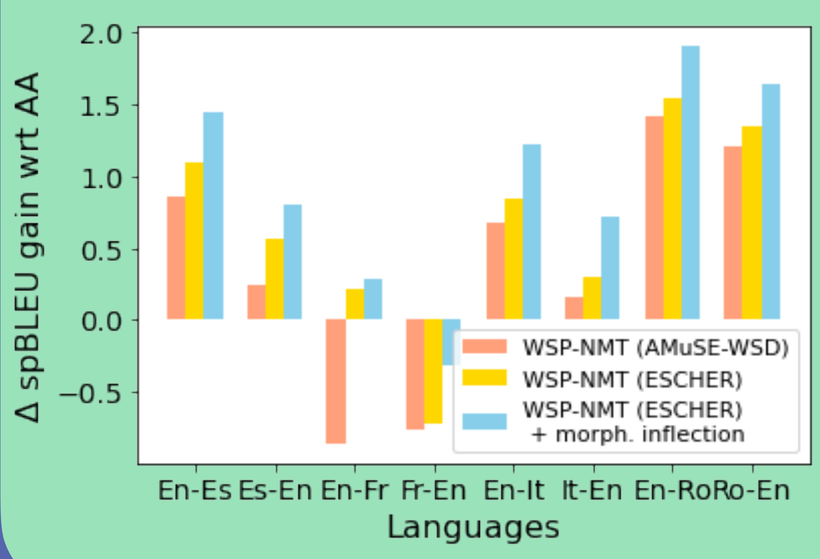
5. APPROACH (WSP-NMT)



a) Baseline: Aligned Augmentation (AA) *(Sense-Agnostic Pretraining)*

b) Our method: WSP-NMT *(Sense-Pivoted Pretraining)*

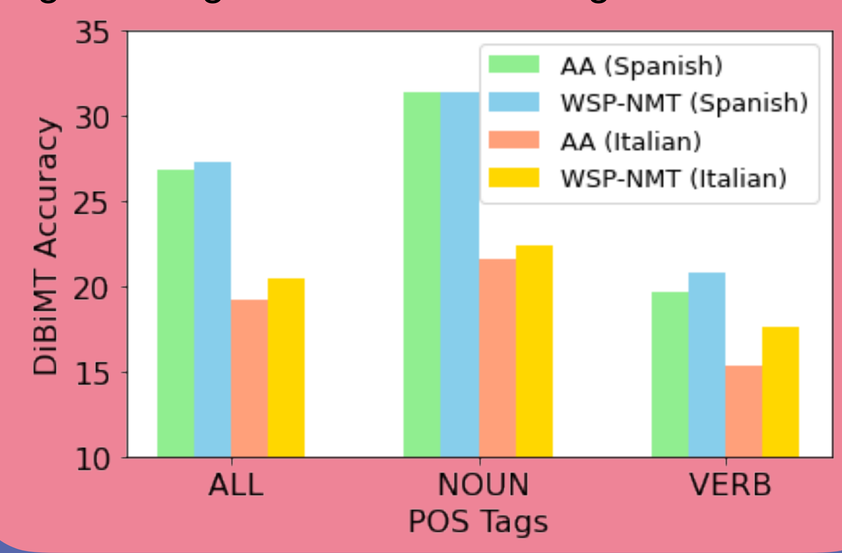
6. RESULTS (OVERALL MT)



- ☑ Consistent gains! Sense-pivoted pretraining helps :-)
- ☑ Better WSD (ESCHER) = better MT quality. But AMuSE-WSD is a good alternative too! (2.3x cheaper)
- ☑ Morph. Inflection Prediction w/ MUSE lexicons for {gender, tense} agreement
- ☑ Lower-resourced En-Ro (5x less data) gains the most!!

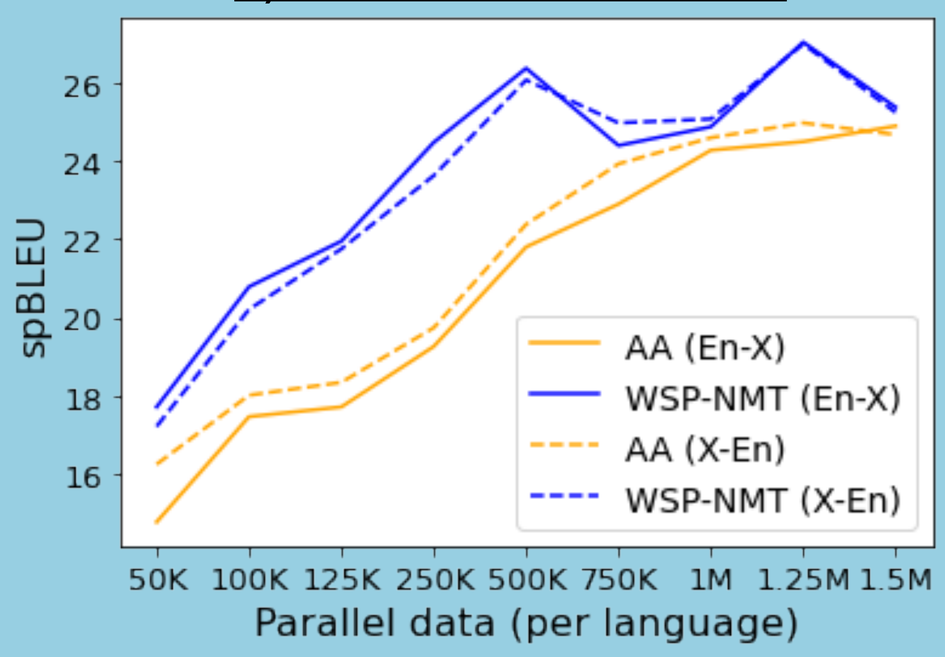
7. RESULTS (AMBIGUOUS MT)

Significant gains in verb disambiguation!



8. SCALING TO RESOURCE-CONSTRAINED SETTINGS

A) Data size vs Performance



Highly effective in low & medium data setups!

B) Zero-Shot Translation

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

Enhanced multilingual convergence!

C) Zero-Shot WSD (Indo-Iranian)

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

Need disambiguation resources :(

9. APPLICATIONS

- ☑ Domain-specific MT
  - less data, well-resourced langs
- ☑ Information-centric domains
  - Healthcare, News etc.

10. CONCLUSION

Advantages:
- ↑ Reliability, ↑ Quality, ↓ Errors
- Useful in low-data setups

Disadvantages:
- Need WSD