Universal Reordering via Linguistic Typology

Joachim Daiber Miloš Stanojević Khalil Sima'an
ILLC, University of Amsterdam

OVERVIEW

Intuition: Linguistic typology describes common and salient features of languages.

Question: Can we exploit typological knowledge for reordering in MT?

Evaluation: Build *universal reordering model* for translation of 22 language pairs.

LINGUISTIC TYPOLOGY

World Atlas of Language Structures [1]

- 2,679 languages
- 192 features
  (32 relevant to reordering)

→ One feature vector per language

EVALUATION

- Phrase-based MT (Moses):
  - k-best WO predictions via lattice
  - No reordering in decoder [2]
- English to many target languages
- Two datasets:
  - Subtitles (800k sentence pairs each)
  - News/Parl. (1m+ sentence pairs each)

MAIN FINDINGS

- Single, universal reordering model
  - built using WALS data
  - works for diverse set of languages
- Linguistic typology
  - can inform reordering models
  - provides adequate descriptions
  - can benefit low-resource setting

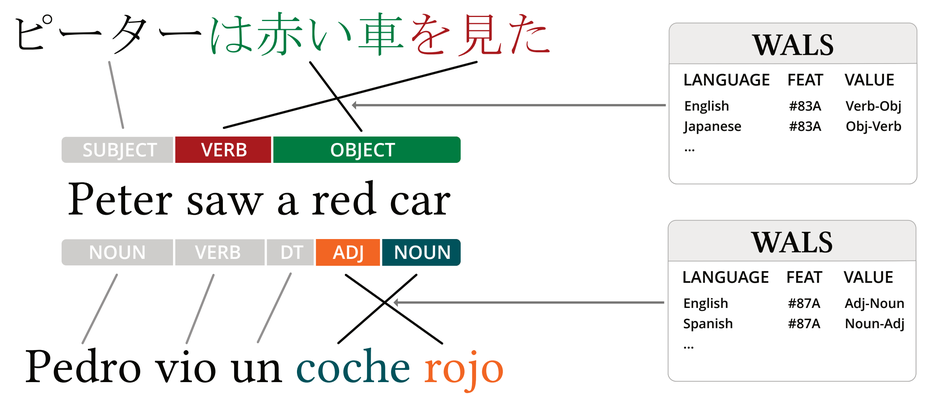
REFERENCES

[1] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

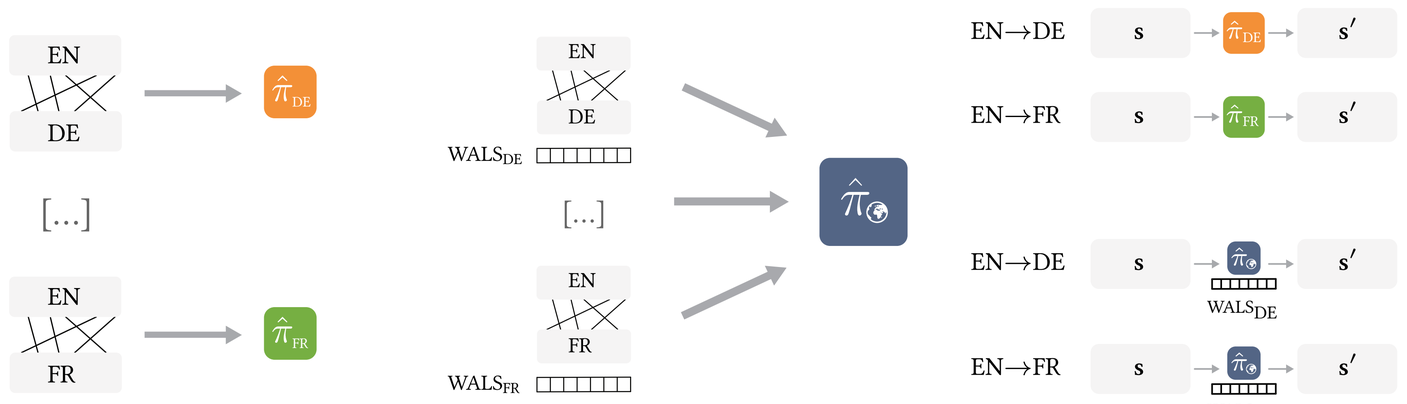
[2] Joachim Daiber, Miloš Stanojević, Wilker Aziz, and Khalil Sima'an. Examining the relationship between preordering and WO freedom in MT. In *WMT 2016*.

[3] Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. Fast and accurate preordering for SMT using neural networks. In *NAACL: HLT 2015*, 2015.

REORDERING & WALS

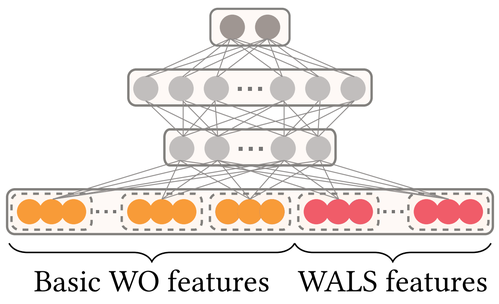


UNIVERSAL REORDERING



(a) Basic preordering training. (b) Universal reord. model training. (c) Applying the models.

BUILDING A UNIVERSAL REORDERING MODEL



- Estimate swap probabilities for source dependency tree nodes
- Graph search for best permutation [3]
- Train combined corpus of all language pairs
- Big variance between language pairs

  → Randomize source WO for class balance

TRANSLATION EXPERIMENTS

Language	BLEU Baseline	Δ BLEU No WALS	Δ BLEU WALS ↓	Δ BLEU Gold	Language	BLEU Baseline	Δ BLEU No WALS	Δ BLEU WALS ↓	Δ BLEU Gold
Dutch	13.76	+0.11	+0.79	+3.44	Greek	7.22	−0.02	+0.01	+0.49
Italian	23.59	+0.04	+0.48	+1.83	Arabic	5.36	−0.10	−0.01	+0.36
Turkish	5.89	−0.36	+0.43	+0.80	Swedish	25.60	−0.14	−0.03	+2.04
Spanish	23.82	−0.27	+0.29	+1.98	Slovenian	10.56	−0.35	−0.10	+1.21
Portuguese	25.94	−0.48	+0.21	+1.64	Slovak	15.56	−0.09	−0.13	+1.98
Finnish	9.95	+0.13	+0.16	+0.51	Icelandic	14.97	−0.31	−0.14	+0.66
Hebrew	11.64	+0.30	+0.11	+2.24	Polish	17.68	−0.45	−0.16	+0.40
Romanian	16.11	+0.11	+0.11	+1.14	Russian	20.12	−0.47	−0.17	+0.92
Hungarian	8.26	−0.10	+0.10	+0.61	German	17.08	−0.21	−0.19	+3.31
Danish	26.36	−0.13	+0.08	+1.56	Czech	12.81	−0.47	−0.21	+0.70
Chinese	11.09	−0.32	+0.05	+0.44	French	19.92	−0.70	−0.23	+1.20

ACKNOWLEDGEMENTS

This work received funding from EXPERT under EU FP7 Marie Curie ITN grant nr. 317471, NWO VICI grant nr. 277-89-002, DatAptor project STW grant nr. 12271 and QT21 project (H2020 nr. 645452).