Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter

Verena Blaschke,[1] Masha Fedzechkina,[2] Maartje ter Hoeve[2]
[1]LMU Munich & MCML (work done while interning at Apple), [2]Apple

Summary

Low-resource evaluation languages with no training data?
→ Cross-lingual transfer!

We carry out cross-lingual transfer experiments between a total of 263 languages (from 33 families) on 3 NLP tasks and analyze how the results correlate with 10 similarity measures.

Different similarity measures matter for different experiments!

Similarity measures

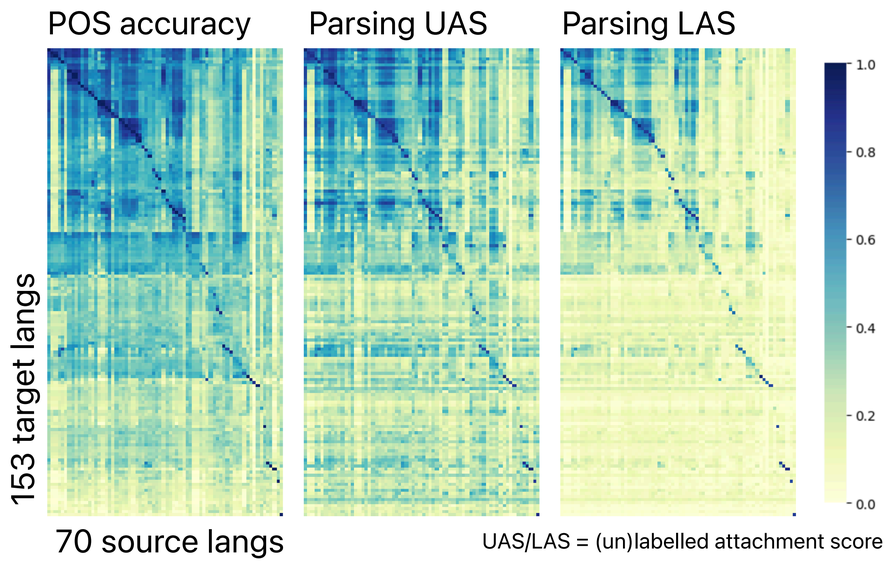
Linguistic measures
- Similarity of **gram**mar, **syn**tax, **pho**nology, phoneme **inv**entory, **lex**icon
- Phylo**gen**etic relatedness
- **Geo**graphic proximity

Dataset measures
- Overlap of **char**acter or **word*** sets
- Training data **size**

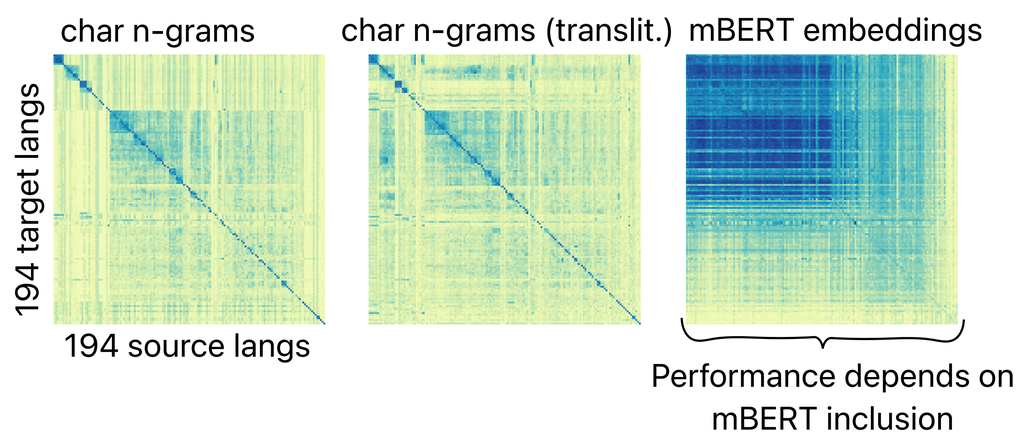
*Depending on the experiment: words, char trigrams, or subwords

Transfer patterns differ across tasks & input representations!

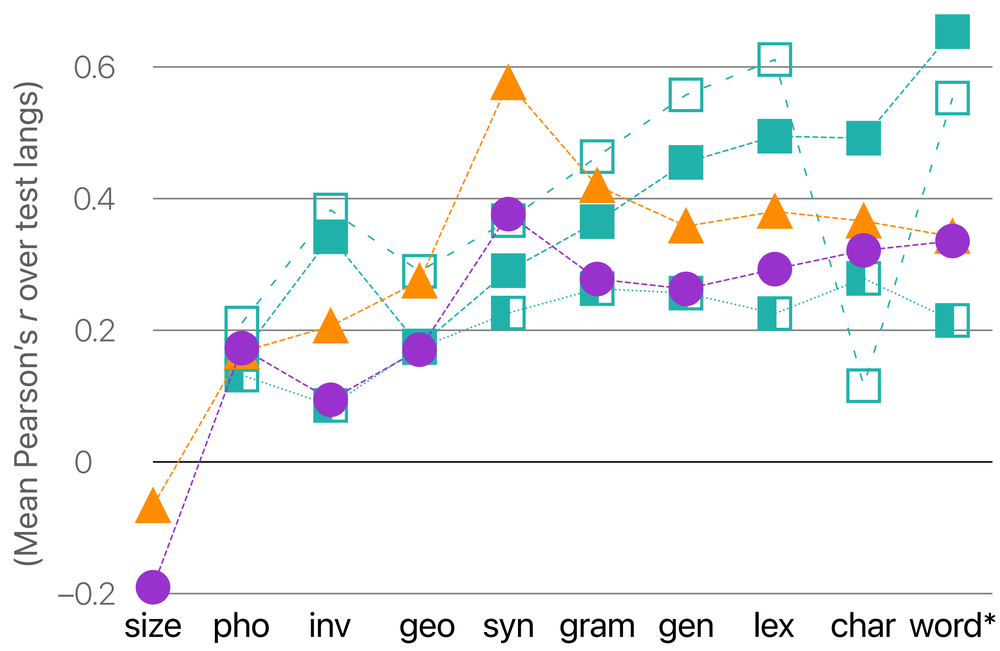
POS tagging & dep. parsing show similar patterns
- Data: Universal Dependencies; model: UDPipe 2 (combines mono- & multilingual embeddings)



Topic classification: transfer patterns differ based on input representations
- Data: SIB-200; metric: accuracy; models: MLPs w/ diff. input representations (lightweight but competitive)



Different experiments = different correlations between task results and similarity measures!



Experiment	Highest correlations
● Parsing (LAS)	syntax
▲ POS	syntax, word overlap
■ Topics (n-grams)	trigram/char overlap, lexicon
□ Topics (n-grams, translit.)	trigram overlap, lexicon
◧ Topics (mBERT)	—

Some trends are shared:
- Uncorrelated: training data size, phonology
- The transfer results can't be predicted by any one factor alone.

Picking a source language for cross-lingual transfer

What if I can't run hundreds of experiments to choose a source language?

Select a source language based on...
- the transfer results for a similar task (w/ similar input representations)
- a similarity measure strongly correlated w/ the task results

Ideally: compare multiple top source language candidates

Mean performance loss (pp.) if picking the source language based *solely* on a given similarity measure / transfer result (darker = worse losses):

	size	pho	inv	geo	syn	gram	gen	lex	char	word*	POS	LAS	UAS	T(n-g)	T(n-g,t)	T(m)
POS	29	15	14	15	10	12	9	10	15	12	—	2	3	9	10	16
LAS	21	13	13	13	7	10	8	8	16	11	1	—	1	11	10	18
UAS	27	16	15	16	8	13	9	10	17	14	3	1	—	12	10	17
Topics (n-grams)	—	17	17	13	15	14	9	9	13	4	10	11	13	—	3	19
Topics (n-grams, translit.)	—	13	13	11	11	10	7	7	20	3	8	8	9	3	—	18
Topics (mBERT)	—	12	11	10	9	8	8	8	12	9	4	5	5	7	6	—

similarity measure → global correlations provide helpful guidance

transfer result