Does manipulating tokenization aid cross-lingual transfer?

A study on POS tagging for non-standardized languages

Verena Blaschke, Hinrich Schütze & Barbara Plank

Center for Information and Language Processing, LMU Munich
Munich Center for Machine Learning
verena.blaschke@cis.lmu.de

The tokenization of Alemannic German is much worse than that of Standard German (Especially tricky since Alemannic has no standardized orthography)

M'r	redd	alemànnischi	Mundàrte .
M, ', r	red, ##d	al, ##em, ##à, ##nn, ##isch, ##i	Mund, ##à, ##rte .
Wir	sprechen	alemannische	Mundarten .
Wir	sprechen	al, #emann, ##ische	Mund, ##arten .
W(r	sprechen	alemaInische	Mundarten .
W, (, r	sprechen	al, ##ema, ##In, ##ische	Mund, ##arten .

"We speak Alemannic dialects", as tokenized by GBERT.

No Alemannic training data to remedy this? **Inject character-level noise** [1] into the Standard German finetuning data:

- Randomly select n % of a sentence's words
- For each of these, either delete a random char, replace a random char, or insert a random char

To what extent does noise injection help?

- POS tagging via zero-shot transfer
- UPOS-tagged data from 3 language families (IE, Afro-Asiatic, Uralic)
- Monolingual BERTs/RoBERTas (source language), mBERT, XLM-R
- Noise levels n: 0 / 15 / 35 / 55 / 75 / 95 %

→ Code incl. tagset conversion scripts at github.com/mainlp/noisydialect

PLM choice matters for the target languages

Performance on unseen dialects is much poorer than on the standardized finetuning languages

Magnitude of performance drops in cross-dialectal set-ups depends on the language

Source	Target	Monolingual PLM						mBERT						XLM-R					
Noise:		0	15	35	55	75	95	0	15	35	55	75	95	0	15	35	55	75	95
German	Alsatian G.	44	71	76	77	78	77	58	76	78	78	77	76	46	71	76	78	77	77
German	Swiss German	55	78	80	80	79	78	62	78	78	79	78	77	56	77	79	79	79	78
German	*German*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*
German	Low Saxon*	18	35	48	51	58	60	36	61	66	68	67	67	26	44	58	71	71	71
Dutch	Low Saxon*	52	62	63	64	64	63	73	75	75	75	73	72	63	71	73	73	73	72
Dutch	*Dutch*	*98*	*97*	*97*	*95*	*93*	*83*	*97*	*97*	*97*	*96*	*95*	*92*	*98*	*98*	*97*	*96*	*96*	*94*
Bokmål	East N.	35	60	67	65	62	60	57	60	58	57	56	54	66	63	63	62	61	59
Bokmål	North N.	36	63	69	67	65	62	61	61	61	60	60	58	70	66	66	65	64	62
Bokmål	West N.	33	59	66	63	61	59	58	57	56	55	54	53	67	62	61	60	59	57
Nynorsk	East N.	64	69	67	65	62	59	59	59	56	56	55	53	67	66	64	62	60	57
Nynorsk	North N.	67	72	69	68	65	63	62	61	59	60	59	57	71	68	67	66	64	62
Nynorsk	West N.	65	69	66	64	63	60	58	58	56	56	56	54	68	64	63	61	60	58
Bokmål	*Bokmål*	*99*	*98*	*98*	*97*	*96*	*91*	*98*	*98*	*97*	*97*	*96*	*92*	*99*	*98*	*98*	*98*	*97*	*93*
Nynorsk	*Nynorsk*	*98*	*98*	*97*	*97*	*95*	*90*	*97*	*97*	*96*	*96*	*94*	*90*	*98*	*97*	*97*	*96*	*95*	*92*
French	Picard	48	52	52	52	51	48	68	73	74	73	73	72	67	74	76	76	75	75
French	*French*	*89*	*88*	*86*	*83*	*78*	*66*	*98*	*98*	*97*	*97*	*96*	*93*	*98*	*98*	*98*	*98*	*97*	*94*
French	Occitan*	41	44	45	45	45	44	86	87	86	85	85	83	77	81	83	83	82	82
Spanish	Occitan*	62	69	70	69	69	69	83	84	83	82	81	79	72	79	78	79	78	77
Spanish	*Spanish*	*99*	*99*	*97*	*97*	*96*	*89*	*99*	*99*	*98*	*96*	*96*	*91*	*99*	*99*	*98*	*98*	*97*	*93*
MSA	Egyptian A.	67	70	66	62	57	50	59	61	60	58	54	47	64	66	65	62	57	50
MSA	Gulf Arabic	66	69	65	61	56	49	65	65	62	60	55	49	66	66	65	61	57	49
MSA	Levantine A.	64	65	62	58	53	47	56	57	55	53	50	45	59	61	60	57	53	46
MSA	Maghrebi A.	51	54	53	50	46	42	50	51	49	48	46	42	51	53	52	50	47	42
MSA	*MSA*	*94*	*93*	*89*	*83*	*78*	*67*	*96*	*95*	*91*	*85*	*79*	*69*	*96*	*95*	*91*	*86*	*80*	*70*
Finnish	Ostroboth. F.	81	80	79	77	78	75	78	78	76	74	73	70	81	85	86	86	86	84
Finnish	SE Finnish	81	79	77	75	76	73	75	75	73	70	69	66	81	84	84	84	84	82
Finnish	SW Finnish	75	73	72	71	71	70	68	68	67	64	63	61	76	80	80	81	81	79
Finnish	SW trans. area	79	78	77	76	76	74	72	72	70	68	67	65	79	84	84	85	84	83
Finnish	Savonian F.	82	80	78	76	76	73	77	79	76	73	72	69	81	84	85	85	85	83
Finnish	Tavastian F.	81	80	79	78	78	75	76	77	76	73	72	69	81	85	86	86	86	84
Finnish	*Finnish*	*98*	*98*	*98*	*97*	*96*	*94*	*96*	*96*	*96*	*95*	*94*	*93*	*98*	*97*	*97*	*97*	*96*	*94*

No universal best noise level!

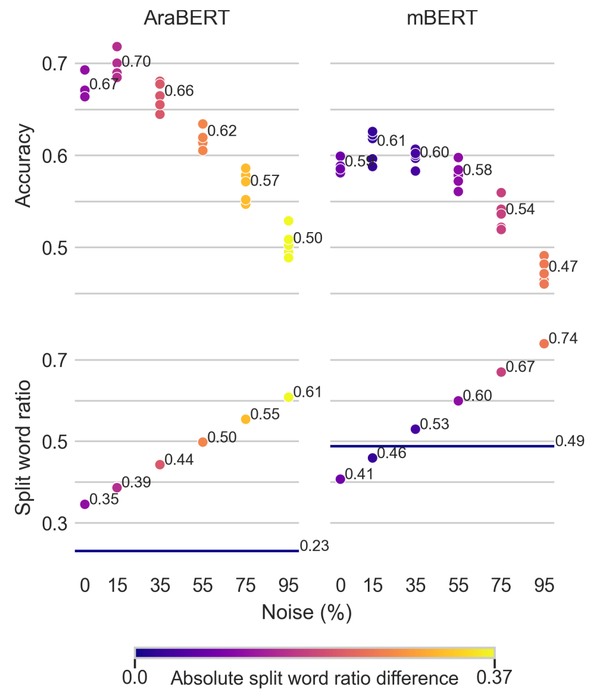
- Usually: 15 % noise is much better than no noise
- Sometimes, *much* greater gains can be made with more noise
- Sometimes, adding any noise at all hurts the performance
- Typically, accuracy as a function of noise has a single global maximum and no local maxima

Accuracy scores (in %), averaged over five initializations. Target languages marked with an asterisk* appear in the training data for mBERT.

Does noise injection make the representation of the finetuning & target data more similar?

→ *Split word ratio difference* correlates best with accuracy: the (absolute) difference between the ratios of words split into subword tokens in the source and target data

The higher the noise level, the larger the *split word ratio* of the source data (→ rising sequences of dots)



Transfer from MSA to Egyptian Arabic. *Top:* Accuracy scores per language model and noise level. *Bottom:* Split word ratios per language model and noise level for the source data (dots) and the target data (dark blue lines).

Generally: the smaller the *split word ratio difference* between source and target (= the darker the dots), the higher the accuracy

- (Strong, but not perfect, correlations)

Recommendation:

- If you don't want to tune noise level as a hyperparameter: compute the *split word ratio differences* for different noise levels & **pick the noise level with the lowest difference**
- Otherwise, search for the best noise level: start low & **increase the noise level until the dev accuracy starts dropping**

Reference

[1] N. Aepli & R. Sennrich (2022). Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. *Findings of the Association for Computational Linguistics.*

This research is supported by the European Research Council (ERC) Consolidator Grant DIALECT 101043235 and in parts Advanced Grant NonSequeToR 740516.