Geographically Grounded Language Models

Valentin Hofmann
Allen Institute for AI

LTL Research Seminar
01/02/2024

How Is This Thing Called?

How Is This Thing Called?

Sofa?

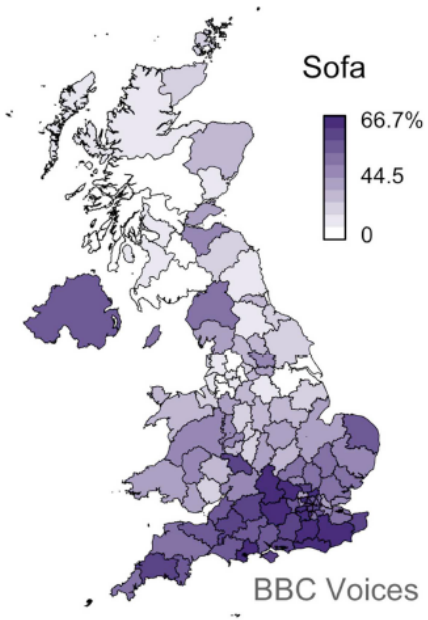
How Is This Thing Called?

Sofa? Couch?

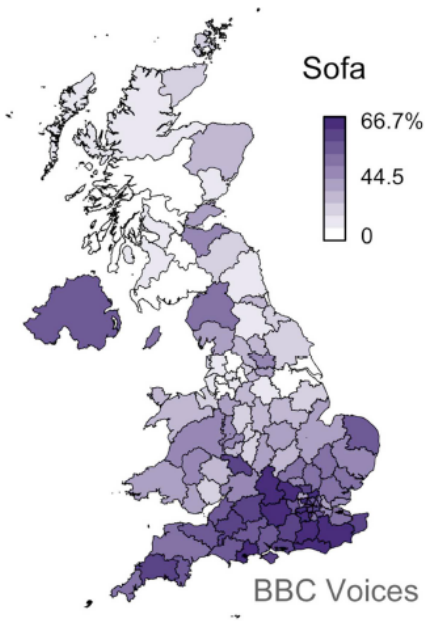
How Is This Thing Called?

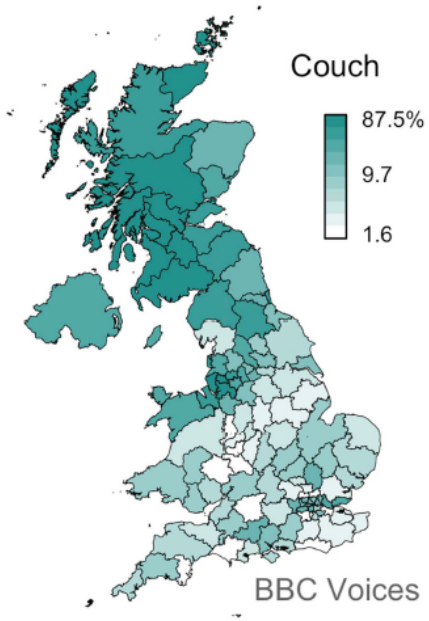
Sofa? Couch? Settee?

It Depends!

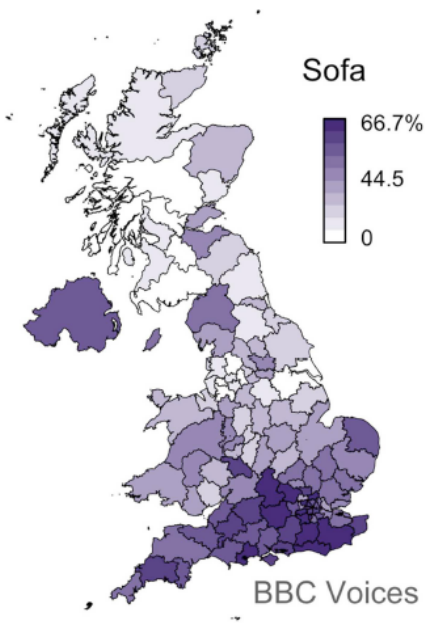


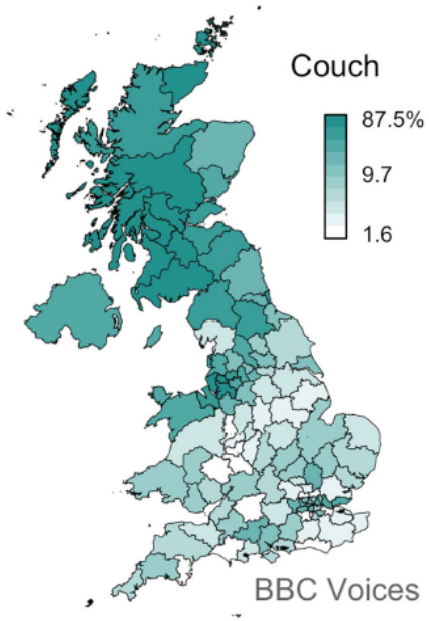
It Depends!

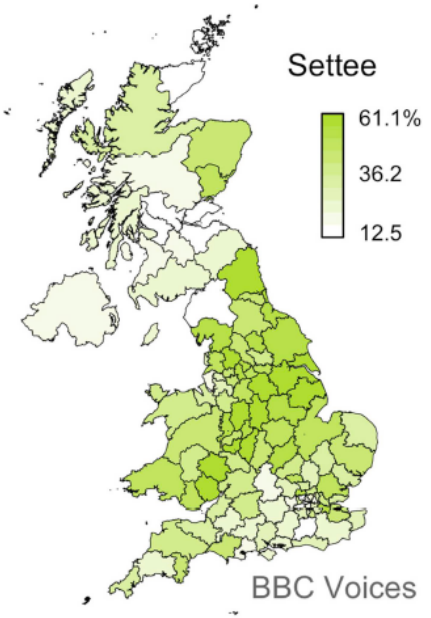




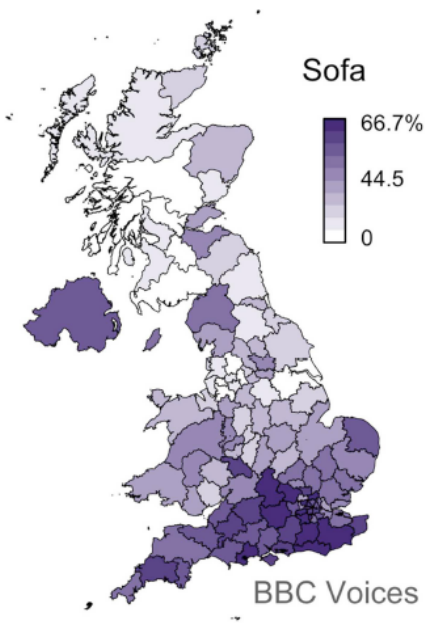
It Depends!

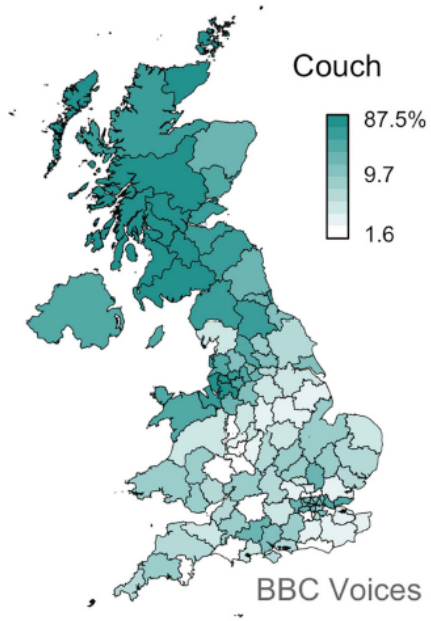

Sofa
66.7%
44.5
0
BBC Voices

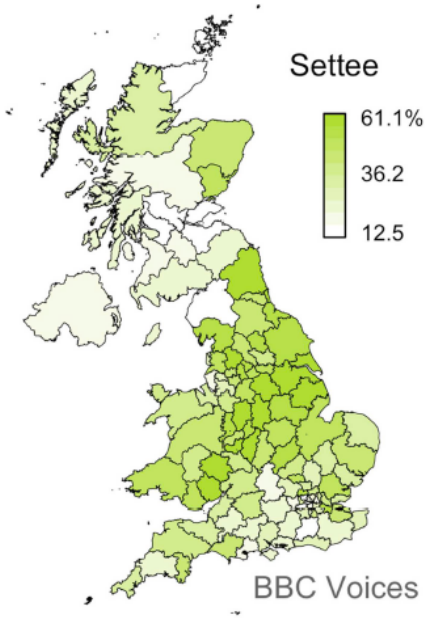

Couch
87.5%
9.7
1.6
BBC Voices


Settee
61.1%
36.2
12.5
BBC Voices

It Depends!

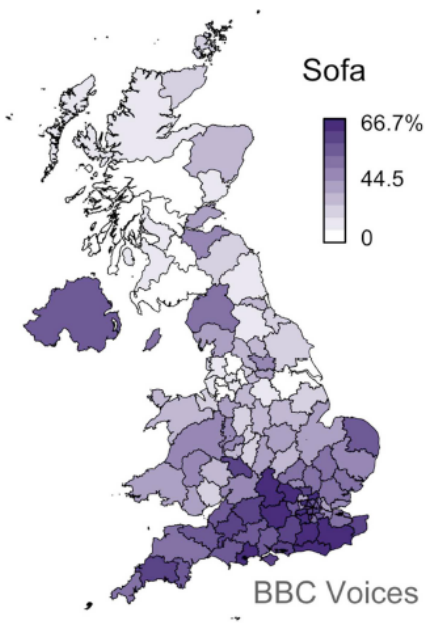


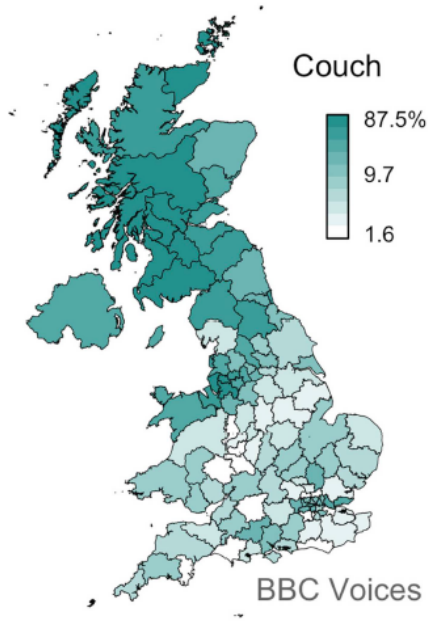


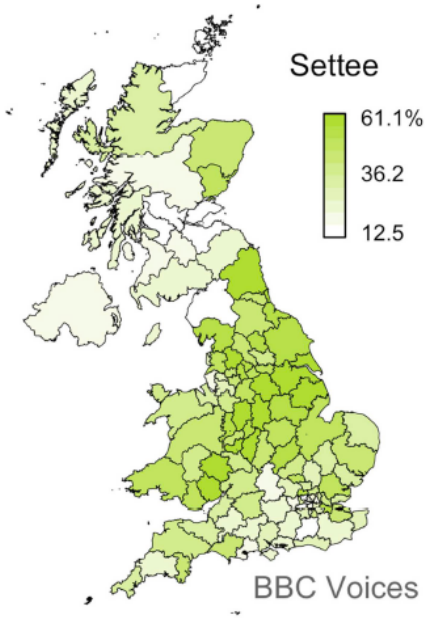


Language varies geographically ...

It Depends!







Language varies geographically ...

... but LMs are only trained on text data!

Key Questions of This Talk

(1) How much knowledge about geographical variation in language do LMs acquire during text-only pretraining?

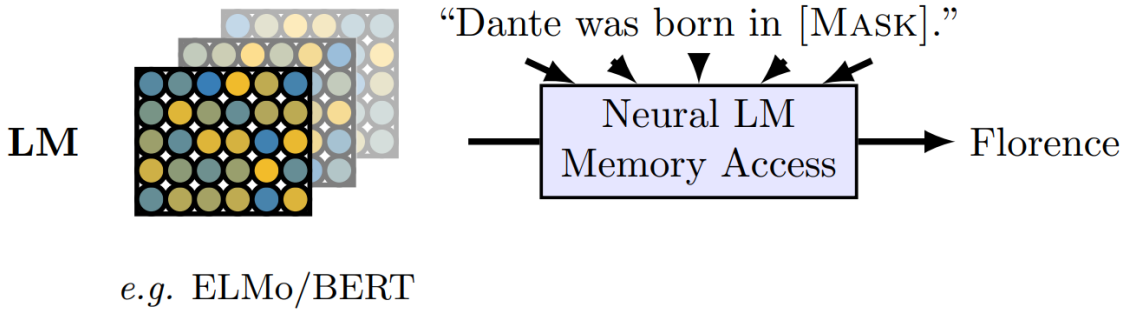
(2) Does additional geographical grounding of LMs enhance their performance on relevant tasks?

Outline

Outline

What Kinds of Knowledge Do LMs Possess?

- Linguistic knowledge (Mahowald et al., 2023)
- Factual knowledge (Petroni et al., 2019)
- Linguistic and factual knowledge of LMs can be accessed in a zero-shot/few-shot manner



Anything Else?

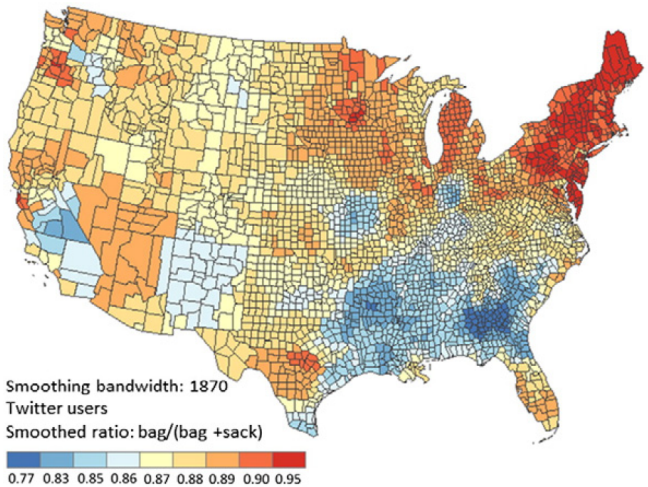
- Humans are known to possess rich sociolinguistic knowledge, which they exploit during language processing
  - Gender (Lass et al., 1979)
  - Ethnicity (Trent, 1995)
  - Geography (Clopper and Pisoni, 2004)
- So far relatively little interest in sociolinguistic knowledge of LMs in NLP (but see, e.g., Lauscher et al., 2022)

Anything Else?

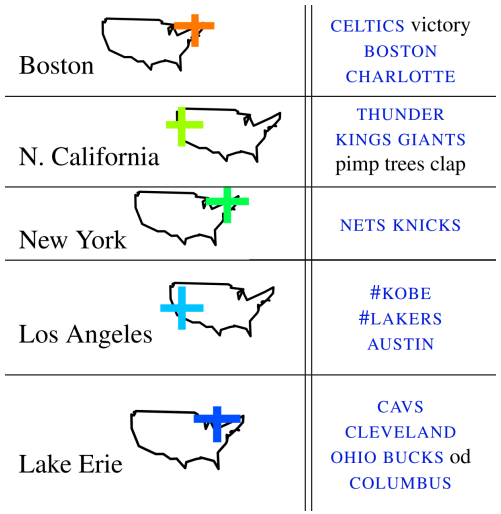
- Humans are known to possess rich sociolinguistic knowledge, which they exploit during language processing
  - Gender (Lass et al., 1979)
  - Ethnicity (Trent, 1995)
  - Geography (Clopper and Pisoni, 2004) ← focus of this talk
- So far relatively little interest in sociolinguistic knowledge of LMs in NLP (but see, e.g., Lauscher et al., 2022)

Geographical Variation in Language

- Main causes: dialectal variation, regional topic variation
- Corresponding knowledge = geolinguistic knowledge



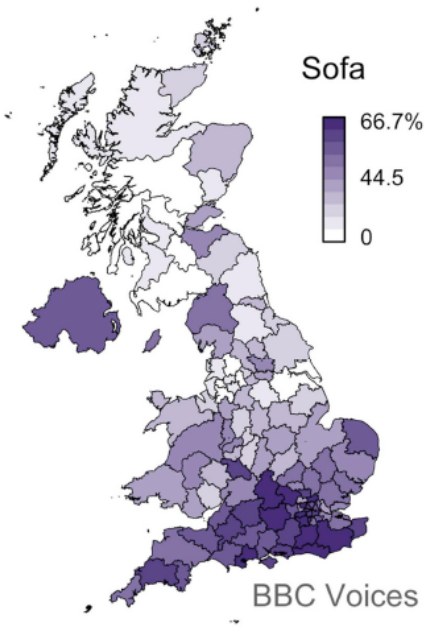
Dialectal variation

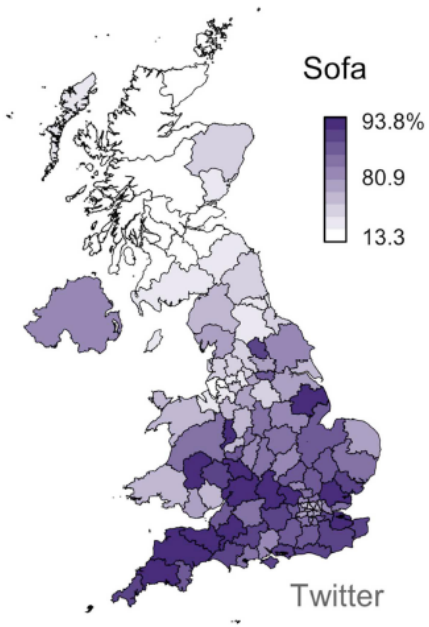


Regional topic variation

Do We Actually Need Geographical Grounding?

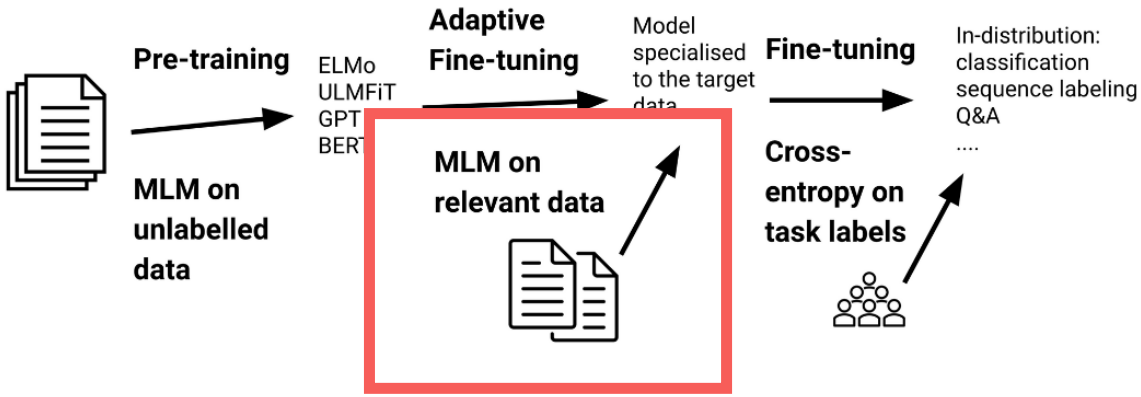
- Social media widely encodes geographical variation in language
- Typical pretraining datasets contain a lot of social media data
- It is unclear whether this is enough for the LMs to acquire geolinguistic knowledge





General Framework

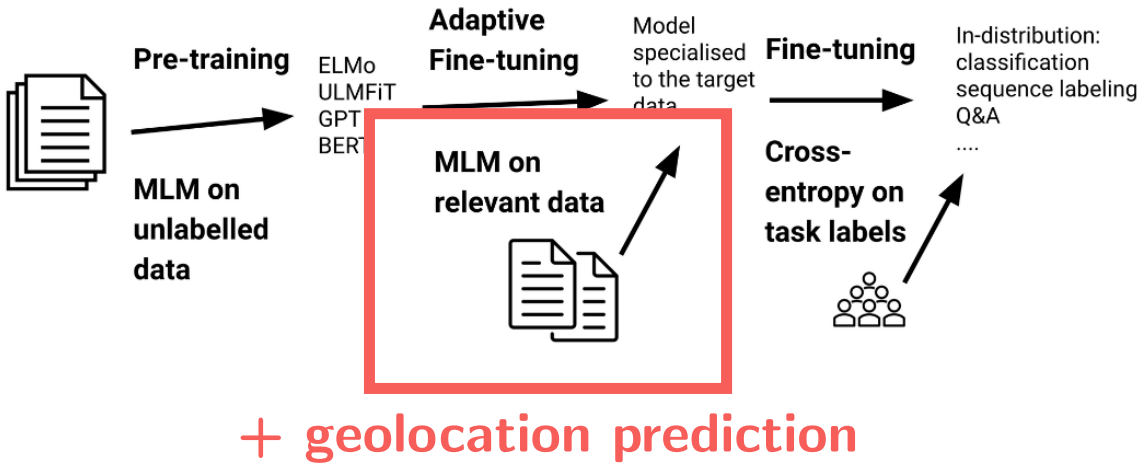
- We conduct continued pretraining of LMs
- Two main settings:
  - Continued pretraining *without* geographical grounding
  - Continued pretraining *with* geographical grounding
- After continued pretraining, we evaluate the geolinguistic knowledge of LMs using different tasks



Outline

How Can We Geographically Ground LMs?

- Objective: encourage LMs to create more direct assocations between linguistic phenomena and geographical locations
- Idea: combine continued pretraining with geolocation prediction in a multi-task learning setup (geoadaptation)
- Requires geotagged text data



Vanilla Geoadaptation

- Forward pass through LM trunk

Vanilla Geoadaptation

- Forward pass through LM trunk
- Token-level output embeddings fed into two heads
  - Language modeling head computes loss $\mathcal{L}_{\text{lm}}$
  - Geolocation prediction head computes loss $\mathcal{L}_{\text{geo}}$ (double regression based on longitude/latitude values)

Vanilla Geoadaptation

- Forward pass through LM trunk
- Token-level output embeddings fed into two heads
  - Language modeling head computes loss $\mathcal{L}_{\text{lm}}$
  - Geolocation prediction head computes loss $\mathcal{L}_{\text{geo}}$ (double regression based on longitude/latitude values)
- Multi-task loss: $\mathcal{L}_{\text{mt}} = \mathcal{L}_{\text{lm}} + \mathcal{L}_{\text{geo}}$

Uncertainty Weighting

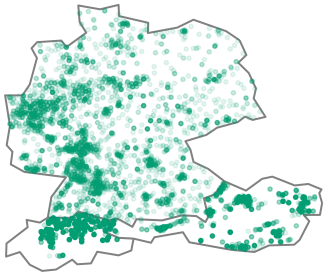
- Simple sum of $\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{geo}}$ might not be ideal
  - $\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{geo}}$ measured on different scales
  - Model confidence might differ for the two tasks

Uncertainty Weighting

- Simple sum of $\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{geo}}$ might not be ideal
  - $\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{geo}}$ measured on different scales
  - Model confidence might differ for the two tasks
- Solution: weight $\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{geo}}$ based on their homoscedastic uncertainties σ_{lm} and σ_{geo} (Kendall and Gal, 2017)
- Multi-task loss: $\mathcal{L}_{\text{mt}} = \frac{1}{2\sigma_{\text{lm}}^2}\mathcal{L}_{\text{lm}} + \log \sigma_{\text{lm}} + \frac{1}{2\sigma_{\text{geo}}^2}\mathcal{L}_{\text{geo}} + \log \sigma_{\text{geo}}$
- σ_{lm} and σ_{geo} are learned as part of model training

Outline

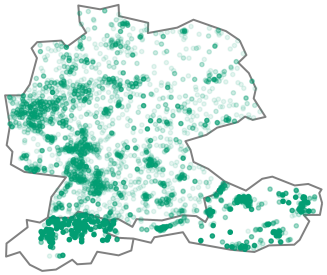
Language areas



AGS

Austrian
German
Swiss

Language areas



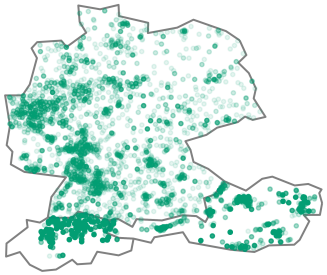
AGS

Austrian
German
Swiss

BCMS

Bosnian
Croatian
Montenegrin
Serbian

Language areas

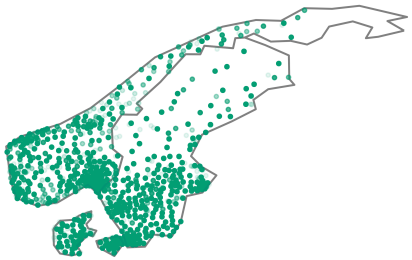


AGS

Austrian
German
Swiss

BCMS

Bosnian
Croatian
Montenegrin
Serbian

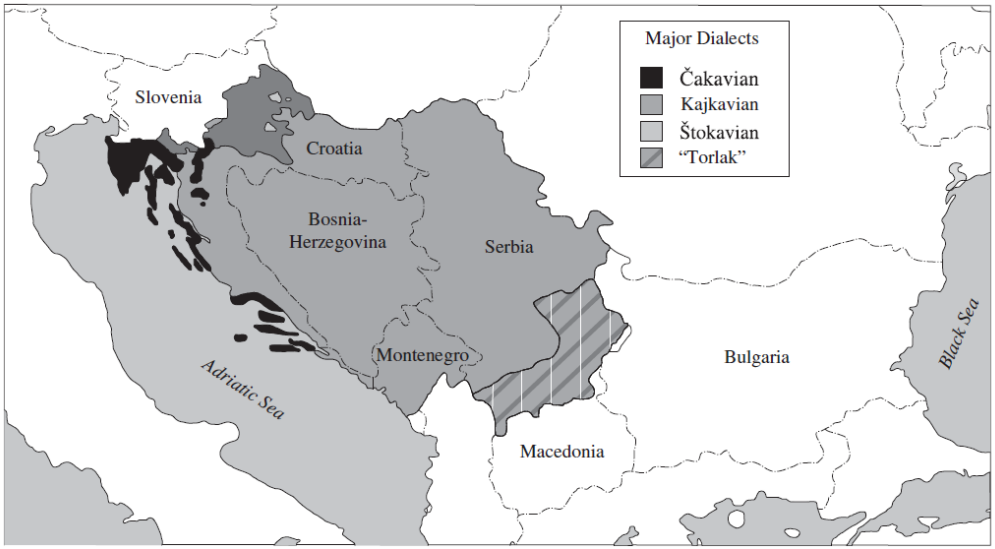


DNS

Danish
Norwegian
Swedish

Excursus: BCMS

- Mutually intelligible South Slavic language(s) spoken in Bosnia-Herzegovina, Croatia, Montenegro, and Serbia
- Serbo-Croatian? Croatian and Serbian? BCS? BCMS?
- Traditional dialect areas based on phonological and lexical variables



BCMS Dialectal Variation

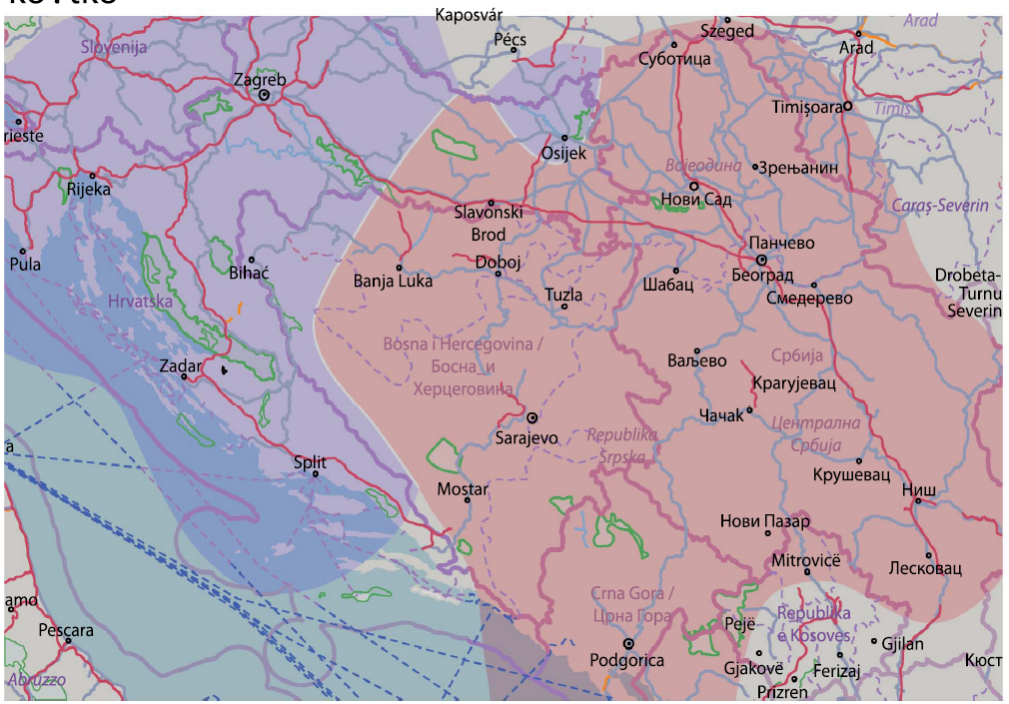
- Systematic phonological/orthographic variation
  - *lijepo* vs. *lepo* ‘nice’
  - *vrijeme* vs. *vreme* ‘time’
- Differences very pronounced on the lexical level
  - *tjedan* vs. *nedelja* ‘week’
  - *vlak* vs. *voz* ‘train’
- Differential dictionaries

BCMS Dialectal Variation

- Systematic phonological/orthographic variation ← perfect for NLP!
  - *lijepo* vs. *lepo* 'nice'
  - *vrijeme* vs. *vreme* 'time'
- Differences very pronounced on the lexical level ← perfect for NLP!
  - *tjedan* vs. *nedelja* 'week'
  - *vlak* vs. *voz* 'train'
- Differential dictionaries ← perfect for NLP!

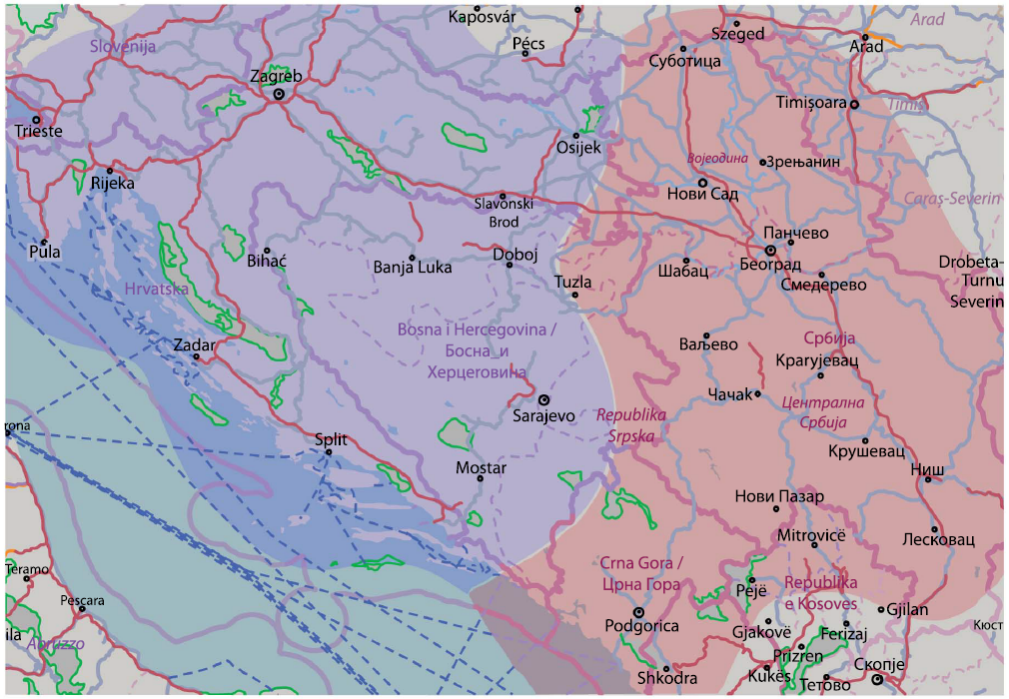
Examples of BCMS Lexical Variation

ko : tko



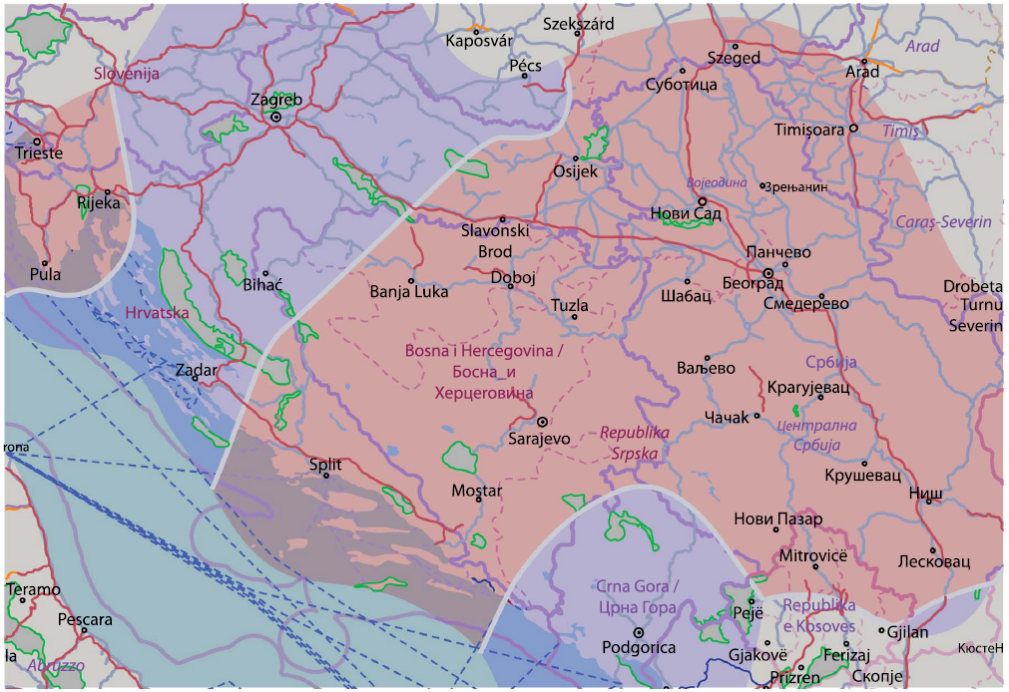
Examples of BCMS Lexical Variation

mnogo : puno

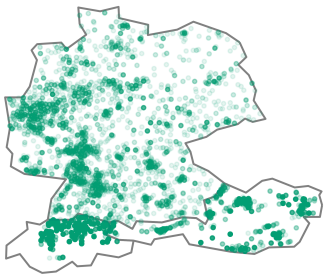


Examples of BCMS Lexical Variation

sta : sto



Models and Data

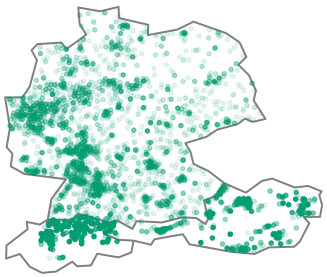


AGS

GermanBERT

Geotagged Jodel posts

Models and Data



AGS

GermanBERT

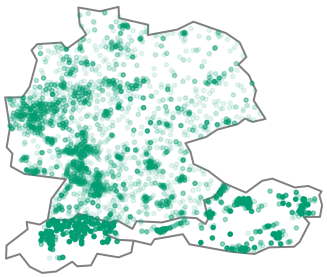
Geotagged Jodel posts

BCMS

BERTić

Geotagged tweets

Models and Data



AGS

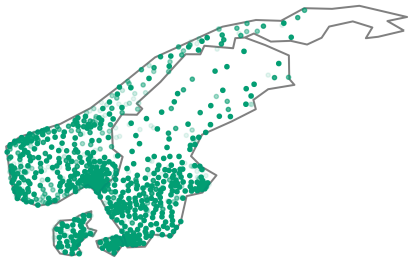
GermanBERT

Geotagged Jodel posts

BCMS

BERTić

Geotagged tweets

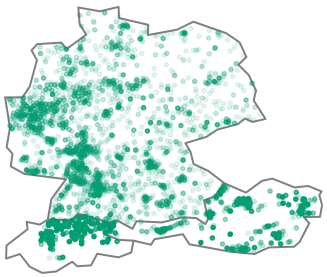


DNS

ScandiBERT

Geotagged tweets

Models and Data

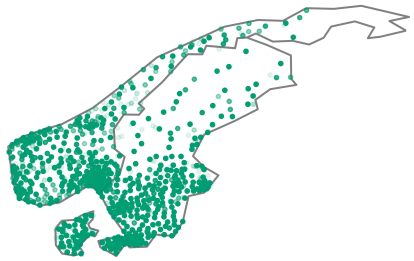


AGS

GermanBERT
Geotagged Jodel posts

BCMS

BERTić
Geotagged tweets



DNS

ScandiBERT
Geotagged tweets

\+ mBERT trained on union of all data (EUR)

How Can We Measure Geolinguistic Knowledge in LMs?

- Direct: zero-shot evaluation
- Indirect: evaluation of performance after task-specific fine-tuning

How Can We Measure Geolinguistic Knowledge in LMs?

- Direct: zero-shot evaluation
- Indirect: evaluation of performance after task-specific fine-tuning
- Higher degree of geolinguistic knowledge should be reflected by performance improvements across various tasks!

Evaluation Tasks

- Fine-tuned geolocation prediction

Evaluation Tasks

- Fine-tuned geolocation prediction
- Fine-tuned language identification

Evaluation Tasks

- Fine-tuned geolocation prediction
- Fine-tuned language identification
- Zero-shot geolocation prediction
  - Add prompt to posts/tweets: *This is [MASK]*
  - Measure probability assigned by LMs to different city names

Evaluation Tasks

- Fine-tuned geolocation prediction
- Fine-tuned language identification
- Zero-shot geolocation prediction
  - Add prompt to posts/tweets: *This is [MASK]*
  - Measure probability assigned by LMs to different city names
- Zero-shot language identification
  - Add prompt to posts/tweets: *This is [MASK]*
  - Measure probability assigned by LMs to different language names

Evaluation Tasks

- Fine-tuned geolocation prediction
- Fine-tuned language identification
- Zero-shot geolocation prediction
  - Add prompt to posts/tweets: *This is [MASK]*
  - Measure probability assigned by LMs to different city names
- Zero-shot language identification
  - Add prompt to posts/tweets: *This is [MASK]*
  - Measure probability assigned by LMs to different language names
- Zero-shot dialect feature prediction (only for BCMS)
  - Measure probability assigned by LMs to phonological variants, e.g., *lepo* (Serbian) vs. *lijepo* (Croatian) 'nice'
  - Measure probability assigned by LMs to lexical variants from differential lexicon, e.g., *voz* (Serbian) vs. *vlak* (Croatian) 'train'

Summary of Methods

- MLMAda: continued pretraining with masked language modeling
- GeoAda-S: geoadaptation with summing of losses
- GeoAda-W: geoadaptation with uncertainty weighting of losses

Summary of Methods

- MLMAda: continued pretraining with masked language modeling
- GeoAda-S: geoadaptation with summing of losses
- GeoAda-W: geoadaptation with uncertainty weighting of losses
- Zero-shot tasks: language modeling head trained in the same way!

Outline

Fine-tuned Geolocation Prediction

- Geoadaptation consistently improves performance
- Uncertainty weighting is better than summing of losses
- Measure: median distance in km

	FT-Geoloc ↓								ZS-Geoloc ↑			
	AGS		BCMS		DNS		EUR					
Method	Dev	Test	Dev	Test	Dev	Test	Dev	Test	AGS	BCMS	DNS	EUR
SotA / Rand	—	—	?30.11	?15.49	—	—	—	—	‡.071	‡.070	‡.026	‡.021
MLMAda	‡193.51	‡196.18	‡29.36	‡16.72	‡101.15	‡101.15	‡107.20	‡107.41	‡.142	‡.144	‡.106	‡.108
GeoAda-S	†190.21	**193.18**	†26.02	†13.98	†98.82	†97.63	†98.00	†101.76	.192	†.287	†.135	†.159
GeoAda-W	**189.06**	†194.85	**23.90**	**12.13**	**95.80**	**97.06**	**97.18**	**97.18**	**.193**	**.319**	**.149**	**.191**

Fine-tuned Language Identification

- Geoadaptation consistently improves performance
- Uncertainty weighting is better than summing of losses
- Measure: accuracy

	FT-Lang ↑								ZS-Lang ↑	
	AGS		BCMS		DNS		EUR			
Method	Dev	Test	Dev	Test	Dev	Test	Dev	Test	BCMS	DNS
Rand	—	—	—	—	—	—	—	—	‡.245	‡.339
GlotLID	—	—	‡.323	‡.316	‡.927	‡.931	—	—	—	—
FastText	‡.843	‡.840	‡.598	‡.588	‡.948	‡.959	‡.757	‡.762	—	—
MLMAda	.851	.855	‡.693	‡.694	‡.964	‡.966	‡.776	‡.777	‡.417	‡.885
GeoAda-S	**.861**	.856	.734	.726	.972	.975	.789	†.786	**.553**	†.896
GeoAda-W	**.861**	**.858**	**.743**	**.734**	**.973**	**.976**	**.792**	**.796**	†.543	**.927**

Zero-shot Geolocation Prediction

- Geoadaptation improves performance massively (e.g., GeoAda-W vs. MLMAda on BCMS: +17.5%)
- Uncertainty weighting is better than summing of losses
- Measure: accuracy

	FT-Geoloc ↓								ZS-Geoloc ↑			
	AGS		BCMS		DNS		EUR					
Method	Dev	Test	Dev	Test	Dev	Test	Dev	Test	AGS	BCMS	DNS	EUR
SotA / Rand	—	—	?30.11	?15.49	—	—	—	—	‡.071	‡.070	‡.026	‡.021
MLMAda	‡193.51	‡196.18	‡29.36	‡16.72	‡101.15	‡101.15	‡107.20	‡107.41	‡.142	‡.144	‡.106	‡.108
GeoAda-S	†190.21	**193.18**	†26.02	†13.98	†98.82	†97.63	†98.00	†101.76	.192	†.287	†.135	†.159
GeoAda-W	**189.06**	†194.85	**23.90**	**12.13**	**95.80**	**97.06**	**97.18**	**97.18**	**.193**	**.319**	**.149**	**.191**

Zero-shot Language Identification

- Geoadaptation consistently improves performance
- Uncertainty weighting and summing of losses perform similarly
- Measure: accuracy

	FT-Lang ↑								ZS-Lang ↑	
	AGS		BCMS		DNS		EUR			
Method	Dev	Test	Dev	Test	Dev	Test	Dev	Test	BCMS	DNS
Rand	—	—	—	—	—	—	—	—	‡.245	‡.339
GlotLID	—	—	‡.323	‡.316	‡.927	‡.931	—	—	—	—
FastText	‡.843	‡.840	‡.598	‡.588	‡.948	‡.959	‡.757	‡.762	—	—
MLMAda	.851	.855	‡.693	‡.694	‡.964	‡.966	‡.776	‡.777	‡.417	‡.885
GeoAda-S	**.861**	.856	.734	.726	.972	.975	.789	†.786	**.553**	†.896
GeoAda-W	**.861**	**.858**	**.743**	**.734**	**.973**	**.976**	**.792**	**.796**	†.543	**.927**

Zero-shot Dialect Feature Prediction

- Geoadaptation consistently improves performance
- Uncertainty weighting and summing of losses perform similarly
- Measure: accuracy

	ZS-Dialect ↑	
Method	Phon	Lex
Rand	‡.501	‡.499
MLMAda	‡.784	‡.872
GeoAda-S	**.870**	.910
GeoAda-W	.858	**.913**

Summary of Results

- LMs acquire a certain degree of geolinguistic knowledge during text-only pretraining
- Geographical grounding substantially improves the geolinguistic knowledge of LMs

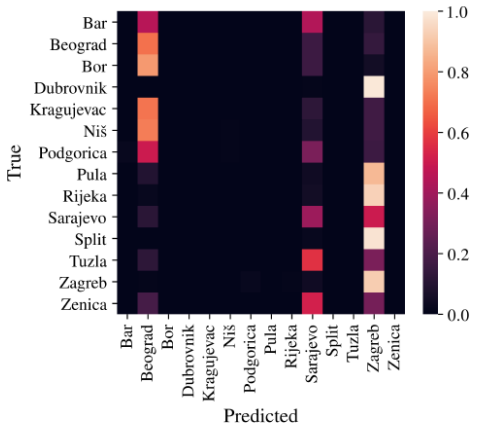
Summary of Results

- LMs acquire a certain degree of geolinguistic knowledge during text-only pretraining
- Geographical grounding substantially improves the geolinguistic knowledge of LMs
- Notice that it is in no way obvious why geoadaptation should help for the zero-shot tasks!

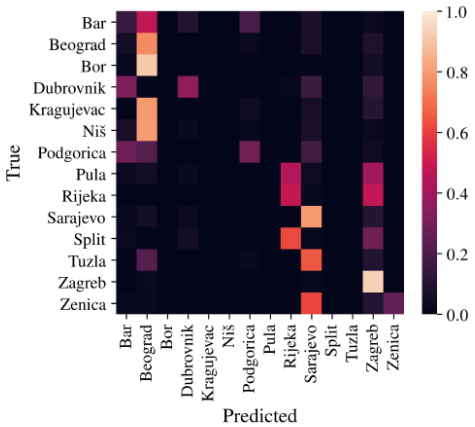
Outline

Where Do the Gains Come From?

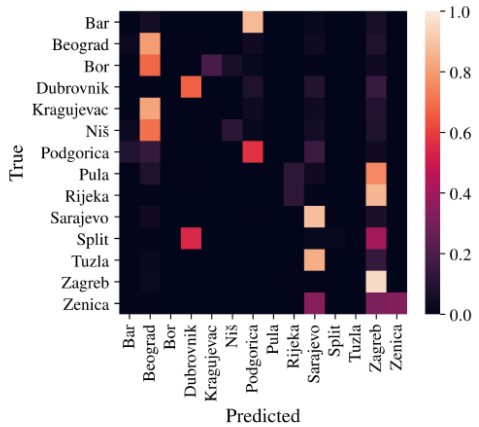
- Analysis of zero-shot geolocation prediction



(a) MLMAda



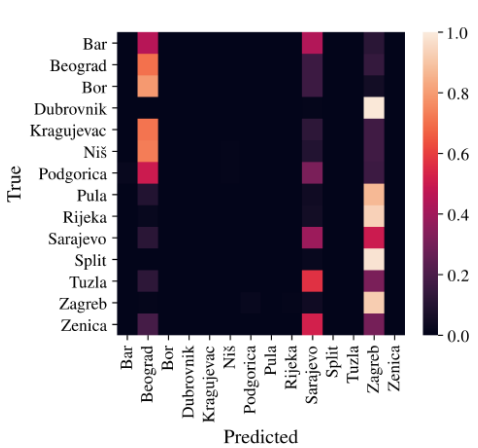
(b) GeoAda-S



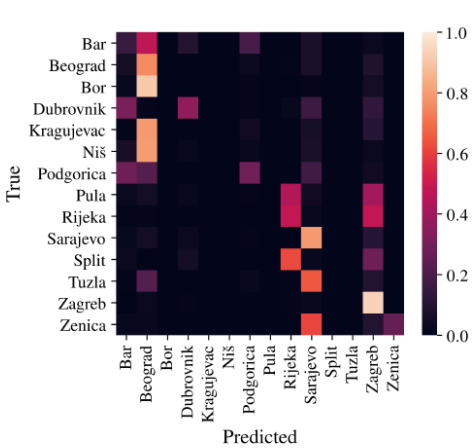
(c) GeoAda-W

Where Do the Gains Come From?

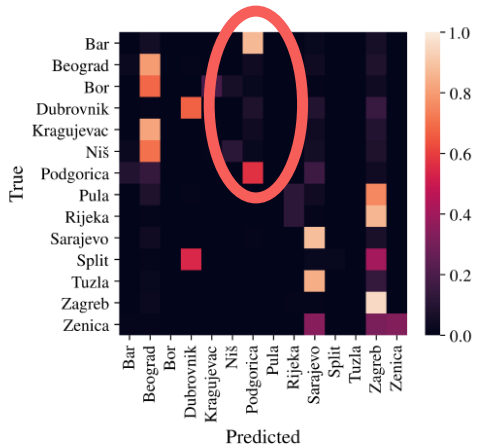
- Analysis of zero-shot geolocation prediction
- Geoadapted models have distinct cluster for Montenegro



(a) MLMAda



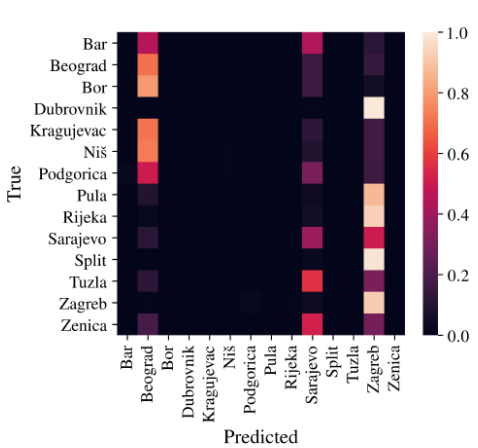
(b) GeoAda-S



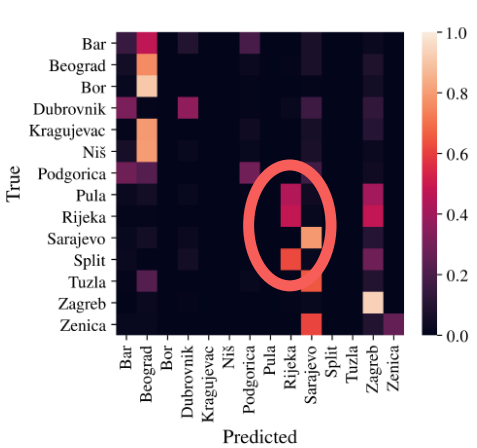
(c) GeoAda-W

Where Do the Gains Come From?

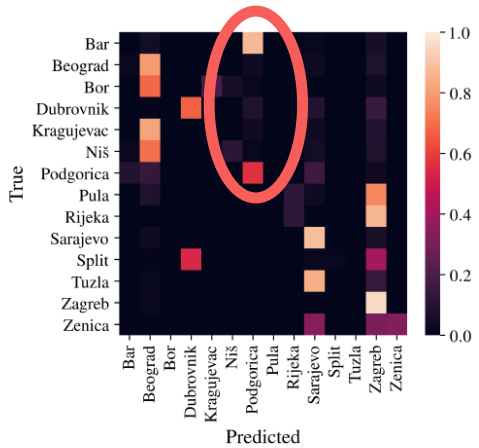
- Analysis of zero-shot geolocation prediction
- Geoadapted models have distinct cluster for Montenegro
- Geoadapted models have distinct cluster for coastal region



(a) MLMAda



(b) GeoAda-S



(c) GeoAda-W

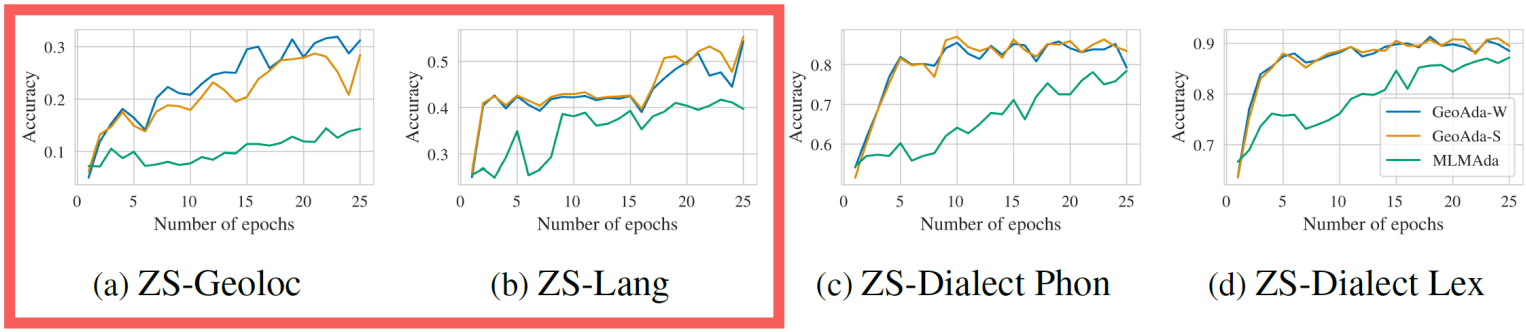
Is It a Problem of Calibration?

- LMs assign different prior probabilities to city names
- Classical solution: calibration (Zhao et al., 2021)
- Calibration helps more for geoadapted models

	ZS-Geoloc ↑			
Method	AGS	BCMS	DNS	EUR
MLMAda	‡.156 ↑.014	‡.150 ↑.006	‡*.131 ↑.025	‡*.139 ↑.031
GeoAda-S	*.229 ↑.036	*.386 ↑.099	.147 ↑.012	†*.195 ↑.036
GeoAda-W	*.229 ↑.036	*.373 ↑.054	.152 ↑.003	*.219 ↑.028

How Does the Geolinguistic Knowledge Build Up?

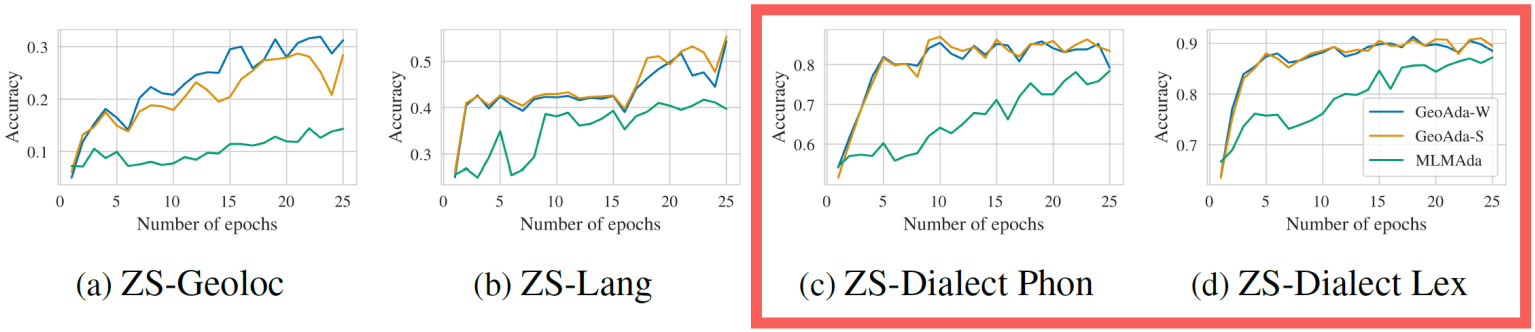
- Geolocation prediction and language identification: MLMAda has almost no advantage through longer adaptation



(a) ZS-Geoloc (b) ZS-Lang (c) ZS-Dialect Phon (d) ZS-Dialect Lex

How Does the Geolinguistic Knowledge Build Up?

- Geolocation prediction and language identification: MLMAda has almost no advantage through longer adaptation
- Dialect feature prediction: geoadaptation (GeoAda-W, GeoAda-S) allows LMs to form geolinguistic knowledge quickly



(a) ZS-Geoloc (b) ZS-Lang (c) ZS-Dialect Phon (d) ZS-Dialect Lex

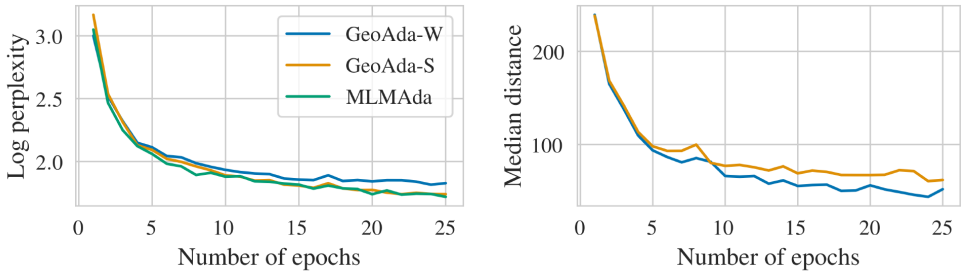
Can We Also Use the CLS Token For Geoadaptation?

- Multi-task geoadaptation based on CLS token (GeoAda-Seq), not sentence tokens as in main experiments (GeoAda-Tok)
- Performs worse except for fine-tuned geolocation prediction

	FT-Geoloc ↓			FT-Lang ↑			ZS-Dialect ↑	
Model	Dev	Test	ZS-Geoloc ↑	Dev	Test	ZS-Lang ↑	Phon	Lex
GeoAda-Seq	†27.35	**12.13**	†.188	.737	.730	†.542	†.844	†.885
GeoAda-Tok	**23.90**	**12.13**	**.319**	**.743**	**.734**	**.553**	**.870**	**.913**

Why Does Uncertainty Weighting Help?

- GeoAda-W has lower geolocation prediction loss at the end of geoadaptation than GeoAda-S
- GeoAda-W weighs geolocation prediction loss more heavily



(a) Language modeling (b) Geolocation prediction

Why Does Geoadaptation Help in the Zero-Shot Setting?

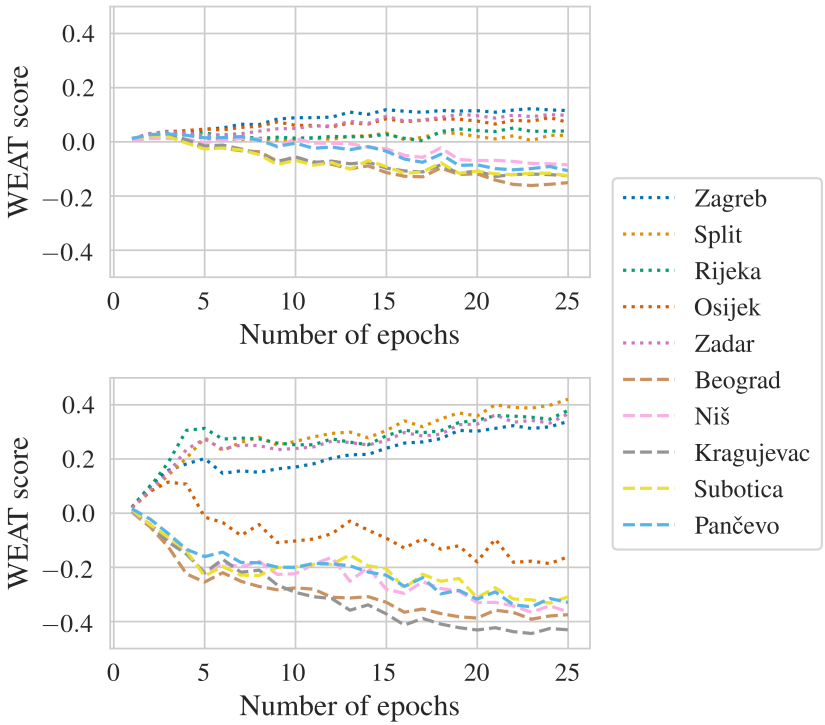
- Hypothesis 1: geographical retrofitting of representations for lexical items that vary geographically (e.g., *lepo* vs. *lijepo*)
  - *Zagreb* in context around masked token causes high probability of *Croatia* for that masked token
  - Training signal from $\mathcal{L}_{\text{geo}}$ should bring *Zagreb* and *lijepo* close to each other in output embedding space
  - After geoadaptation *lijepo* also causes high probability of *Croatia*

Why Does Geoadaptation Help in the Zero-Shot Setting?

- Hypothesis 1: geographical retrofitting of representations for lexical items that vary geographically (e.g., *lepo* vs. *lijepo*)
  - *Zagreb* in context around masked token causes high probability of *Croatia* for that masked token
  - Training signal from $\mathcal{L}_{\text{geo}}$ should bring *Zagreb* and *lijepo* close to each other in output embedding space
  - After geoadaptation *lijepo* also causes high probability of *Croatia*
- Hypothesis 2: geographical retrofitting of representations for toponyms (e.g., *Zagreb*)

Hypothesis 1: Results

- WEAT score of *e*/*ije* variants and Croatian/Serbian cities
- Hypothesis 1 confirmed: associations stronger in the case of geoadaptation compared to vanilla adaptation



Hypothesis 2: Results

- Compare distances of toponym output embeddings (cities) with geodesic distances
- Hypothesis 2 confirmed: correlation higher in the case of geoadaptation (0.872) compared to vanilla adaptation (0.515)

Outline

Key Take-Away Points

- LMs acquire some geolinguistic knowledge during text-only pretraining, but it can be improved using geoadaptation
- The specific form of geoadaptation we propose uses token-level multi-task learning with uncertainty weighting
- Geoadaptation works by geographically retrofitting the token representations of LMs

Thank you!

Massive shout-out to Goran Glavaš, Nikola Ljubešić, Janet Pierrehumbert, and Hinrich Schütze!