

Multilingual NLP

6. Cross-Lingual Transfer

(+ Multilingual Evaluation)

Prof. Dr. Goran Glavaš
Center for AI and Data Science (CAIDAS), Uni Würzburg

After this lecture, you'll...

- Know what cross-lingual transfer for NLP tasks is
- Learn about Massively Multilingual LMs and CL transfer with them
- Distinguish between zero-shot and few-shot CL transfer
- Know of prominent multilingual evaluation benchmarks

Content

- **Cross-Lingual Transfer**
- CL Transfer with Massively Multilingual Transformers (MMTs)
- Zero- and Few-Shot Transfer with MMTs
- Multilingual Evaluation

Why Multilingual NLP?

- **Cross-Lingual transfer:** transfer supervised models for concrete NLP tasks
 - Models trained on labeled data in high-resource source language...
 - ...make predictions on texts in low-resource target languages with little or no labeled data

English



Cross-Lingual Transfer: Practical Necessity

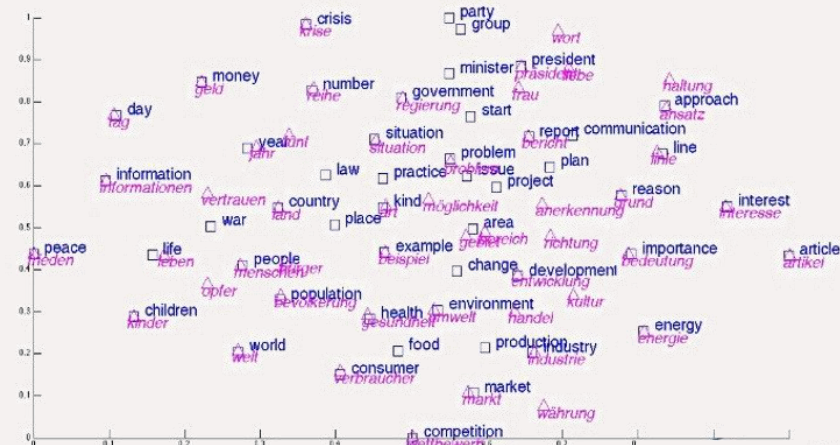
- Only a handful of NLP tasks have annotated data in many languages
 - Part-of-speech tagging (Universal Dependencies, UD)
 - Syntactic parsing (UD)
 - Named Entity Recognition (e.g., WikiANN)
- Higher-level semantic tasks often have only English training data
 - Generally more difficult tasks, e.g.:
 - Natural Language Inference (NLI)
 - Semantic Text Similarity (STS)
 - Question Answering (QA)
 - Causal Commonsense Reasoning
 - ...

Cross-Lingual Transfer: Practical Necessity

- Natural Language Inference
 - Given a **premise** and **hypothesis**, predict whether hypothesis is entailed by the premise, contradicts it, or neither
 - Premise:** „A man reads the paper in a bar with green lighting.“
 - Hypothesis:** „The man is inside“
 - Label:** entailment
- Causal Commonense Reasoning
 - Given a **premise** find its most plausible cause among several **choices**
 - Premise:** „The politician won the election“
 - Choice 1:** „No one voted for him“
 - Choice 2:** „He ran negative campaign ads against the opponent“
 - Label:** Choice 2
- Such **language understanding** datasets are very expensive to build
 - Thus most often exist only in English
 - **Q:** Can't we automatically translate them with MT?

Cross-Lingual Transfer

- **Multilingual representation spaces** necessary for cross-lingual transfer
 - Words/sentences/texts that have the **same/similar meaning**, get same/similar representations...
 - ...whether from the same language or different languages
- Cross-lingual word embeddings
- Multilingual LMs



A decorative graphic consisting of several concentric dotted circles. Three solid black dots are placed at the outer edge of the circles, positioned at approximately the 12, 3, and 9 o'clock positions.

- 



Content

- Cross-Lingual Transfer
- **CL Transfer with Massively Multilingual Transformers**
- Zero- and Few-Shot Transfer with MMTs
- Multilingual Evaluation

Massively Multilingual Transformers

- With pretrained Transformer-based LMs (i.e., BERT & co.)
 - We obtain more than static word embeddings
 - Contextualized representations of tokens - meaning in context
- If we could make the same Transformer (same parameters) learn how to contextualize tokens in multiple languages...
 - We could support CL transfer „out of the box”
 - Fixing for limitations of transfer with CLWEs

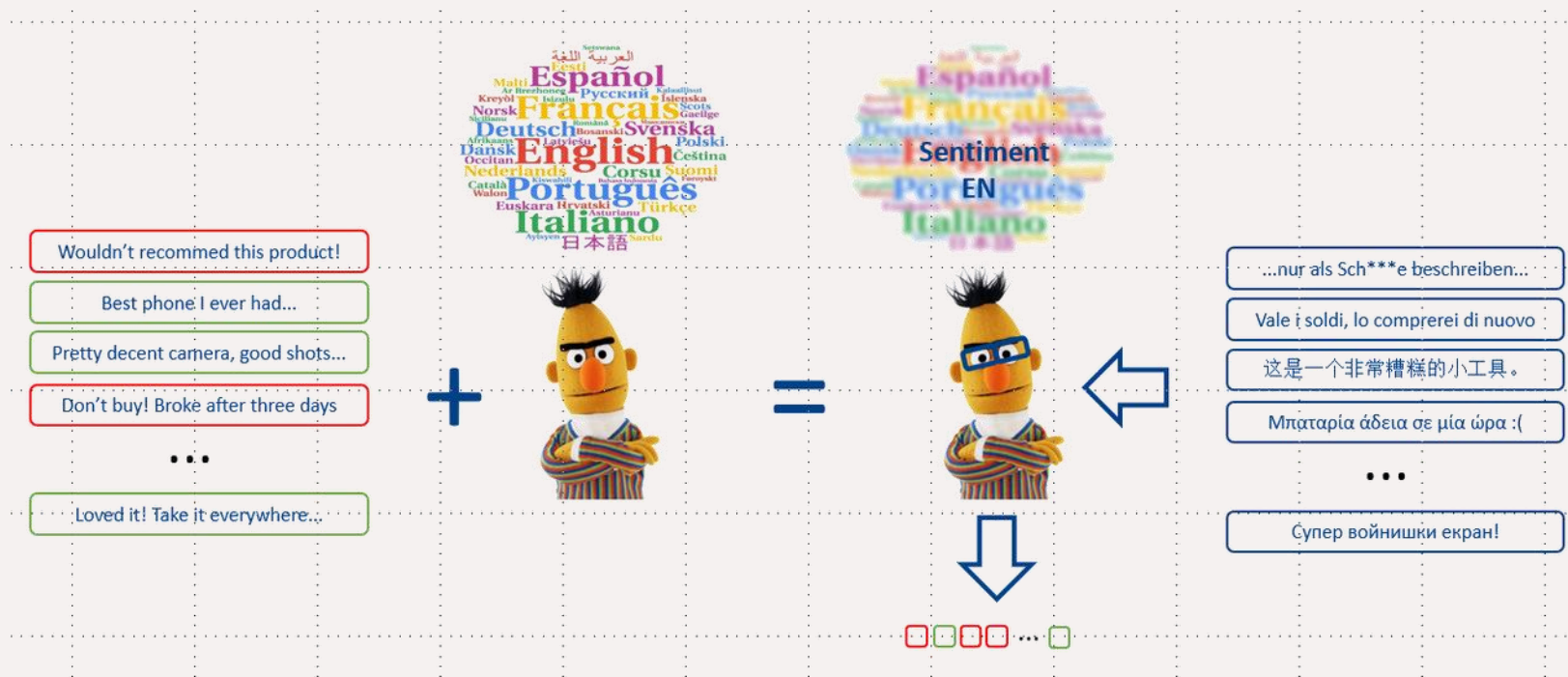
• Multilingual BERT

- BERT's Transformer pretrained on multilingual corpora
- Concatenation of monolingual corpora in 104 languages
- Without any cross-lingual supervision?!
 - No word alignments, no parallel sentences



Massively Multilingual Transformers

- **Cross-lingual transfer with MMTs** is conceptually simple
 1. Place a task-specific head on top of the Transformer body
 2. Perform standard fine-tuning using task-specific training data in L_S
 3. Use the Transformer and classifier to make predictions for data in L_T





Massively Multilingual Transformers

- **Cross-lingual transfer with MMTs** is conceptually trivial
- But a lot of open questions about what's encoded in such an MMT
 - Q: Size of pretraining corpora for each language?
 - Q: How does tokenization work in a massively multilingual setup?
 - Q: How/why are representations of different languages semantically aligned if there is no explicit cross-lingual supervision?
 - Q: Are all pretraining languages „equal“ in the representation space of mBERT?
 - Q: Is CL transfer equally good for any L_S and L_T from pretraining languages?
 - Q: What about languages not seen in pretraining?





MMTs: Corpora and Tokenization

- mBERT trained on 104 largest Wikipedias
 - Obviously, the corpus of each language is not of the same size
 - English Wikipedia: 6.6M articles; Chuvash Wikipedia: 50K articles
 - Articles also much longer for English and other major languages
- Multilingual tokenization
 - mBERT (like monolingual BERT) uses WordPiece tokenization
 - Vocabulary size: 110K tokens
 - Languages without whitespaces:
 - Characters separated with a special character (CJK Unicode block)
 - Problem: WordPiece merges dominated by large languages
 - Large languages have many more whole-word tokens than small languages



MMTs: Corpora and Tokenization

- Problem: WordPiece merges dominated by large languages
 - Large languages have many more whole-word tokens than small languages

```
from transformers import BertTokenizer, BertModel
tokenizer = BertTokenizer.from_pretrained('bert-base-multilingual-uncased')

encoded_input = tokenizer("wonderful", return_tensors='pt')
tokenizer.convert_ids_to_tokens(encoded_input["input_ids"][0])
```

- „wonderful“ (EN) → [' [CLS] ', 'wonderful', ' [SEP] ']
- „prekrasno“ (HR) → [' [CLS] ', 'pre', '##kra', '##sno', ' [SEP] ']

MMTs: Corpora and Tokenization

- Problem: WordPiece merges dominated by large languages
 - Large languages have many more whole-word tokens than small languages
 - „wonderful“ (EN) → [' [CLS]', 'wonderful', '[SEP]']
 - „prekrasno“ (HR) → [' [CLS]', 'pre', '##kra', '##sno', '[SEP]']
- Several shortcomings:
 1. Token sequences longer for smaller languages and Transformer has fixed input size → we can effectively encode shorter texts in smaller languages
 2. We need Transformer's body parameters to correctly contextualize subword tokens that belong to the same word-level token
 - Learn that 'pre', '##kra', and '##sno' should attend over one another
 - But smaller languages have less data to learn from!
 3. Shorter tokens more likely to appear across multiple languages
 - wonderful will appear predominantly in English text, what about ##kra?
 - Shared tokens will commonly have different „meaning“ in different langs

CL Transfer with mBERT



Pires, T., Schlinger, E., & Garrette, D. (2019, July). [How Multilingual is Multilingual BERT?](#) In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4996-5001).

→ „mBERT surprisingly good at zero-shot CL model transfer“



Wu, S., & Dredze, M. (2019, November). [Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT.](#) In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 833-844).

→ „Surprising cross-lingual effectiveness of BERT“

CL Transfer with mBERT

- Q: But where does the cross-lingual transfer ability of mBERT come from?
 - No explicit alignment across languages of any type in pretraining



Dufter, P., & Schütze, H. (2020). Identifying Necessary Elements for BERT's Multilinguality. In Proceedings of EMNLP 2021.

- The capacity of the model (12-layer Transformer; 110M parameters) is too small to precisely and accurately „learn“ every of 104 languages
- MLM training on massively multilingual corpora forces the Transformer to use its parameters efficiently -- i.e., share them across languages
 - This exploits commonalities between languages and results in (some) alignment
- Shared embeddings also help
 - Positional embeddings*: Q: when could shared PEs hurt?
 - Token embeddings, for tokens with same meaning across languages
 - E.g., digits or names („1“, „Joe“, ...)

MMTs beyond mBERT

- XLM: Cross-Lingual Language Modeling



Conneau, A., & Lample, G. (2019). [Cross-lingual language model pretraining](#). Advances in neural information processing systems, 32.

- BPE tokenizer trained on modified corpora obtained by
 - Oversampling sentences from small languages
 - Undersampling sentences from large languages

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha}$$

modified distribution

$$p_i = \frac{n_i}{\sum_{k=1}^N n_k}$$

original distribution

- Smoothing factor α set to 0.5
- More whole-word tokens for small languages, 95K tokens in total

MMTs beyond mBERT

- XLM: Cross-Lingual Language Modeling



Conneau, A., & Lample, G. (2019). [Cross-lingual language model pretraining](#). Advances in neural information processing systems, 32.

- MLM as the main training objective (across all languages)
 - Self-supervised objective
- Additionally leverages parallel data with the new objective named translation language modeling (TLM)
 - Just MLM, but on pairs of parallel sentences
 - TLM is a supervised objective: requires parallel data
 - Also introduces trainable language embeddings

MMTs beyond mBERT



Conneau, A., & Lample, G. (2019). [Cross-lingual language model pretraining](#). Advances in neural information processing systems, 32.

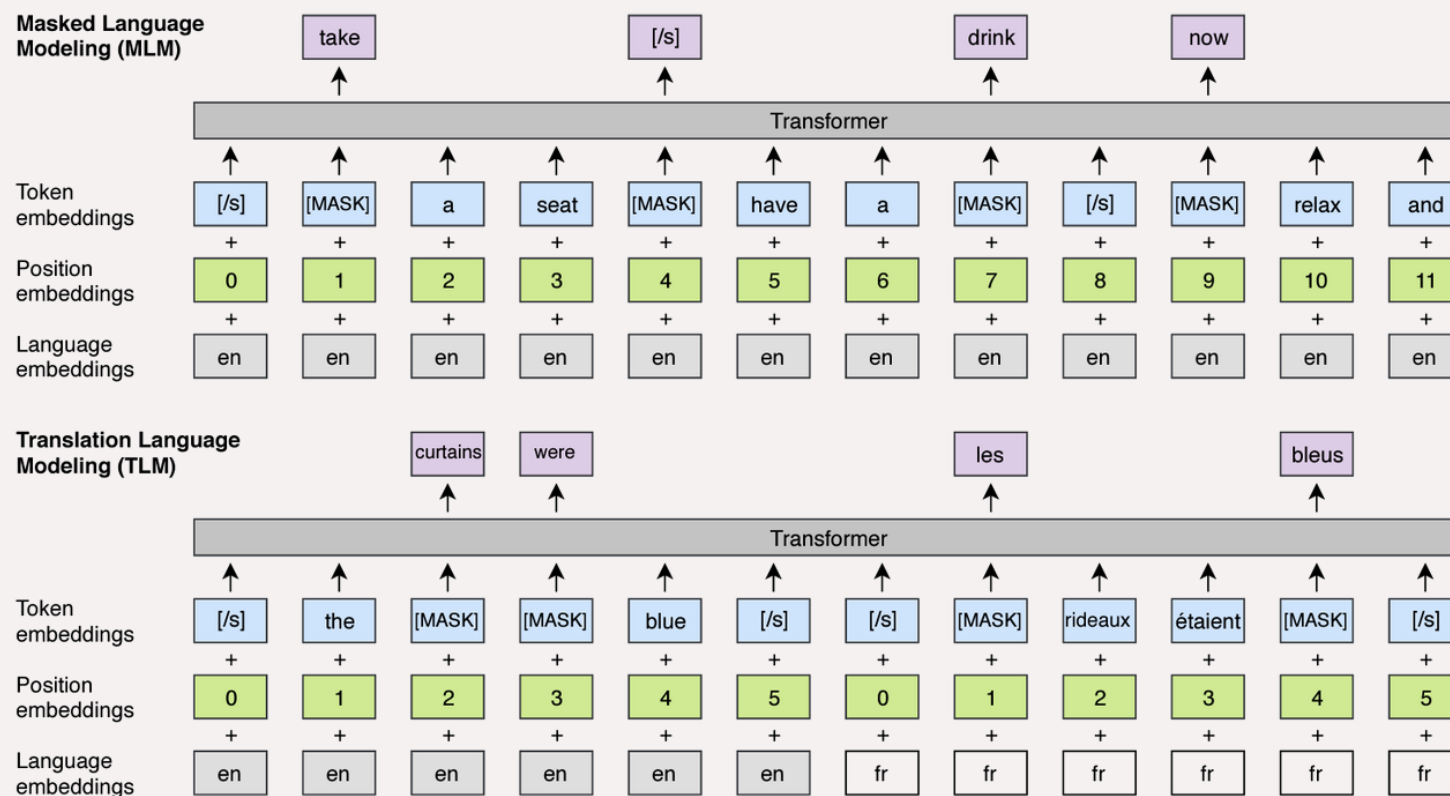


Image from the original paper

MMTs beyond mBERT

- XLM-R: XLM-on-RoBERTa



Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). [Unsupervised Cross-lingual Representation Learning at Scale](#). In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451).

- Just MLM-ing, but...
- On much much larger corpora:
 - [CC100](#) – filtered CommonCrawl for 100 languages: 2TB of text!
- Larger vocabulary: 250K tokens

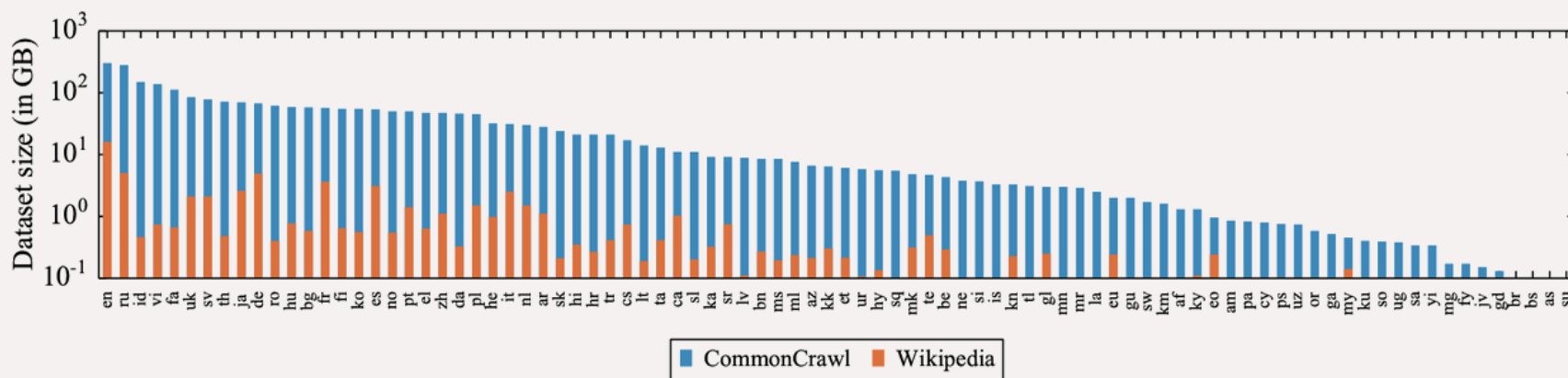


Image from the original paper

CL Transfer with MMTs

- Initial evaluations
 - Source language: EN
 - Target languages: high-resource, closely related to EN
 - E.g., NL, DE, IT, FR, ES
- What about small target languages distant from English?
 - small: small corpus in pretraining
 - distant from English:
 - genealogically, etymologically, typologically (recall Lecture 1 :))
 - Basically, what about the vast majority of world languages?

CL Transfer with MMTs



Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020, November). [From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers](#). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4483-4499).

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	-60.1	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	-85.4	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	-46.1	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	-77.0	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	-31.7	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	-34.8	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	-33.0	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	-20.2	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	-43.2	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	-26.2	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

- **Huge** performance drops (both with mBERT and XLM-R) from transfer to
 - (1) small languages
 - (2) languages distant from English



Poor CL Transfer with MMTs

- MMTs (mBERT, XLM-R) exhibit huge performance drops in CL transfer to low-resource languages, especially if they are distant from English
- Even for large and closely-related languages (e.g., DE, ES, IT) we see drop in performance compared to English.
 - Q: Why?
- For English, we get better results by fine-tuning monolingual English BERT/RoBERTa than by fine-tuning mBERT or XLM-R.
 - Q: Why?



Poor CL Transfer with MMTs

- Part of the problem is the **curse of multilinguality** (Lecture 7)
 - Loss of representational accuracy for each individual language due to representing too many languages with the model of fixed capacity
- MLM training doesn't really align the representations across languages very well: **clusters of language-specific subspaces** visible
 - Better alignment achievable post-hoc with parallel data

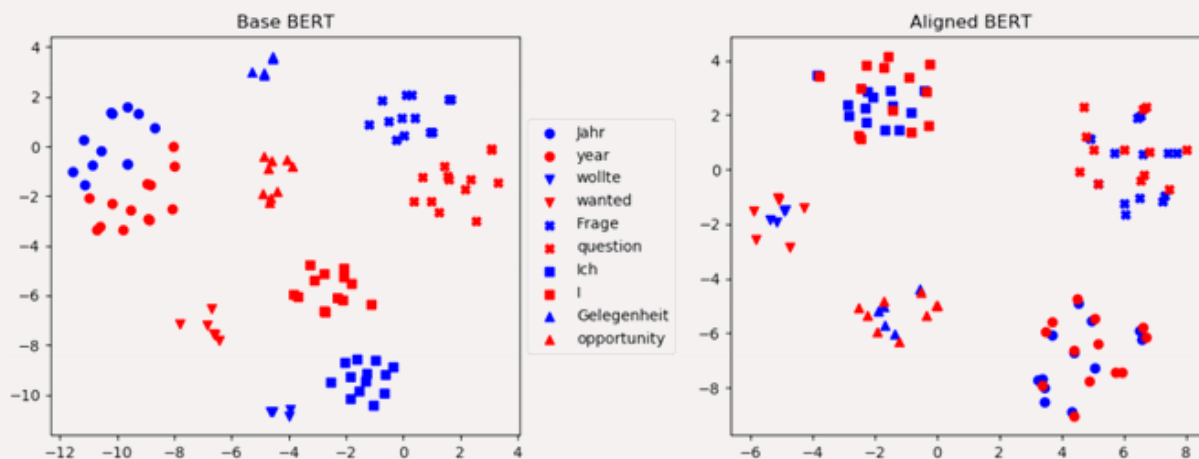


Image from: Cao, S., Kitaev, N., & Klein, D.
[*Multilingual Alignment of Contextual Word Representations*](#). In *International Conference on Learning Representations*. 2020.

Content

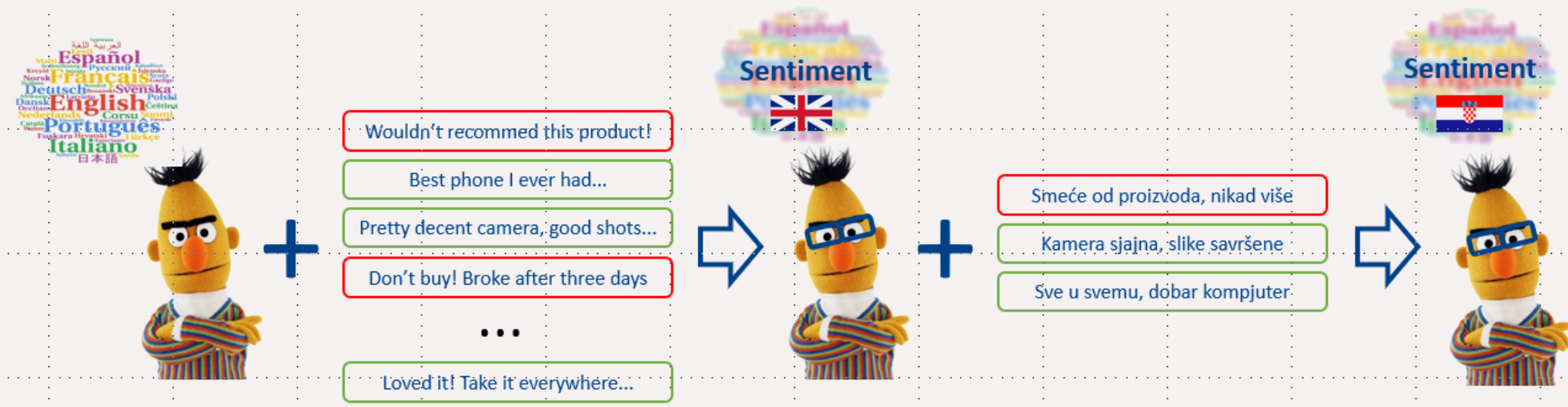
- Cross-Lingual Transfer
- CL Transfer with Massively Multilingual Transformers
- **Zero- and Few-Shot Transfer with MMTs**
- Multilingual Evaluation

Zero- vs. Few-Shot CL Transfer

- So far, we have analyzed the so-called **zero-shot transfer** setup
 - We assume **zero** labeled task instances in the target language
- In practice, it is almost always possible to annotate some small number of instances in the target language
- **Few-shot transfer**: large task-specific training dataset D_S in L_S , plus a few labeled instances (small dataset D_T) in L_T
 - Q: how many is „few“?
 - Depends on the task, but $|D_T| \ll |D_S|$

Few-Shot CL Transfer

- Sequential few-shot CL transfer
 - First fine-tune an MMT on the large D_S
 - Then fine-tune it on the small D_T



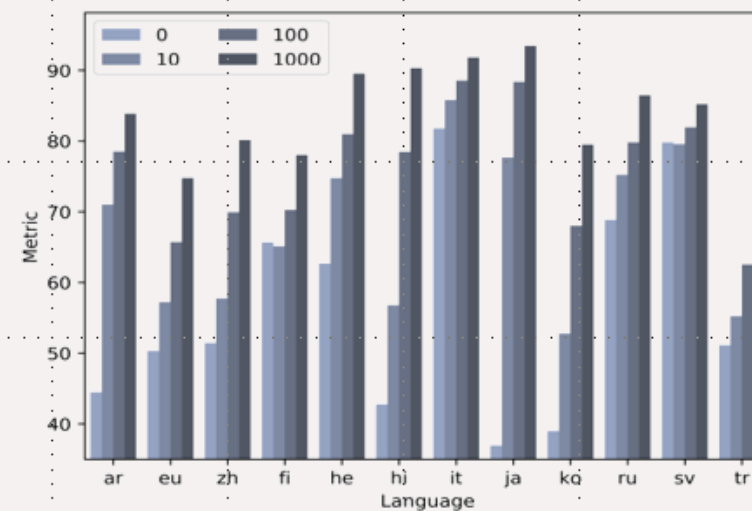
Few-Shot CL Transfer



Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020, November). [From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers](#). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4483-4499).

- Sequential few-shot CL transfer can bring massive gains in transfer performance compared to zero-shot CL transfer

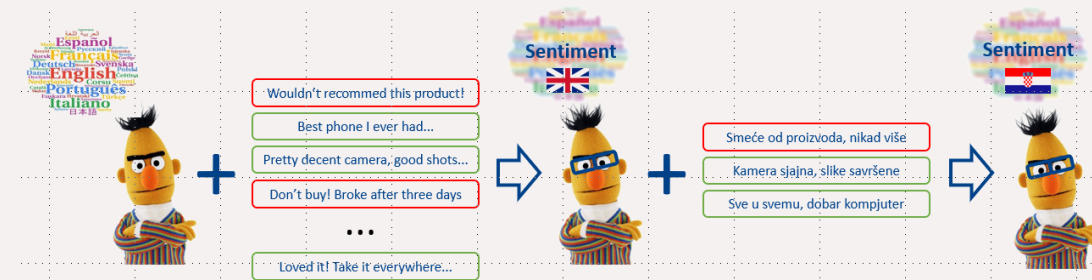
Task	Model	k	$k = 10$		$k = 50$	
		$k = 0$	score	Δ	score	Δ
DEP	mBERT	52.96	66.69	13.73	72.67	19.70
	XLM-R	48.60	65.57	16.97	72.19	23.59
POS	mBERT	67.2	80.17	12.96	85.34	18.14
	XLM-R	65.5	80.68	15.18	85.7	20.2
NER	mBERT	79.34	83.18	3.84	84.54	5.20
	XLM-R	85.43	88.06	2.63	91.07	5.64



Dependency parsing

Few-Shot Transfer: Generality vs. Performance

- Sequential few-shot CL transfer
 - (1) First fine-tune an MMT on the large D_S : computationally **expensive**
 - (2) Then fine-tune it on the small D_T : computationally **cheap**
- **Pro**: After (1) we have a **general** task-specific model, which can be quickly fine-tuned for various target languages with few instances
- **Con**: The two training steps are executed sequentially, there is **no** direct task-specific interaction between the languages

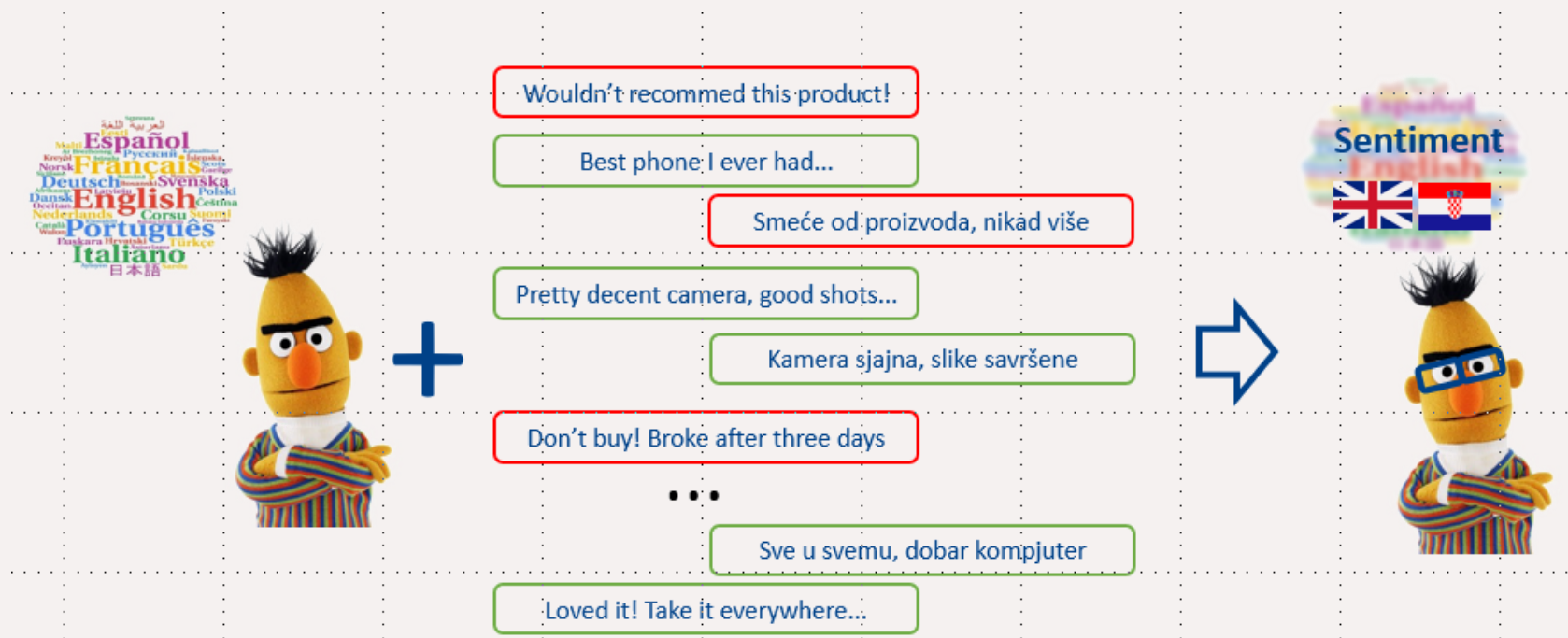


Few-Shot Transfer: Generality vs. Performance



Schmidt, F. D., Vulić, I., & Glavaš, G. (2022). [Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models](#). In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 10725-10742).

- Simultaneous fine-tuning on (many) instances from L_S and (few) from L_T



Few-Shot Transfer: Generality vs. Performance

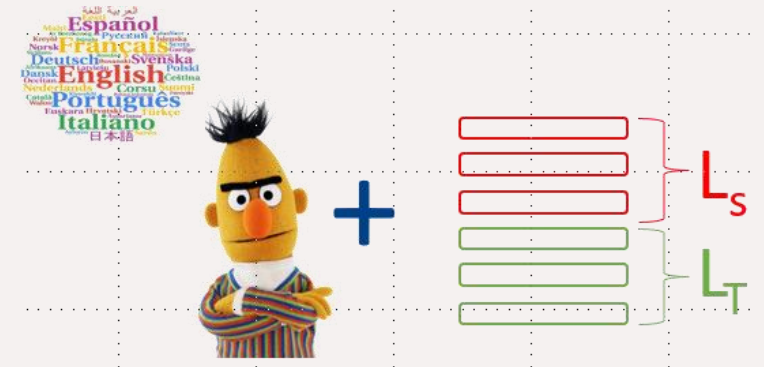


Schmidt, F. D., Vulić, I., & Glavaš, G. (2022). [Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models](#). In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 10725-10742).

- Simultaneous fine-tuning on (many) instances from L_S and (few) from L_T

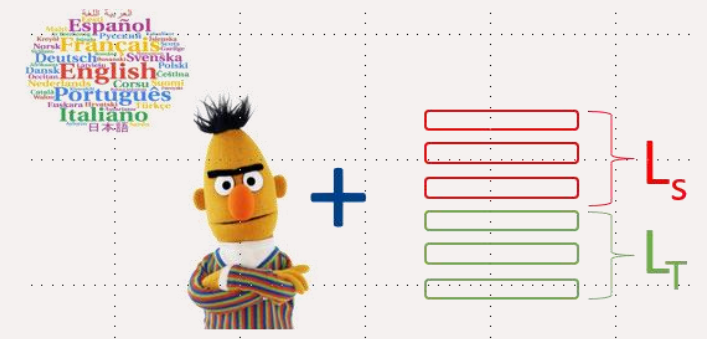
Important: **batch balancing**
between L_S and L_T

- Few target language instances will repeat much more often
- But will be „regularized“ with different source language instances in different batches
→ less overfitting, better generalization in L_T



Few-Shot Transfer: Generality vs. Performance

- Joint few-shot CL transfer
- **Pro:** Task-specific interaction between L_S and L_T , leads to **better performance** on L_T
- **Con:** For each L_T we have to carry out the large fine-tuning on $|D_S| + |D_T|$ instances
 - Effectively $2*|D_S|$ instances in training
 - Because we're repeating D_T instances to balance batches



Content

- Cross-Lingual Transfer
- CL Transfer with Massively Multilingual Transformers
- Zero- and Few-Shot Transfer with MMTs
- **Multilingual Evaluation**



Multilingual Evaluation

- In the last few years, a lot of **new multilingual** evaluation datasets and benchmarks in NLP
- Some multilingual datasets (single task)
 - Americas NLI: evaluation dataset for natural language inference (NLI), covering 10 low-resource indigenous languages of the Americas
 - MasakhaNER: evaluation dataset for named entity recognition (NER) covering 10 low-resource African languages
 - TyDiQA: question answering (QA) dataset covering 11 typologically diverse languages
 - XCOPA: causal commonsense reasoning for 11 genealogically, geographically, and typologically diverse languages



Multilingual Evaluation

- Q: How to select languages for a multilingual dataset/benchmark?
 - Based on what criteria?
- Historically, multilingual evaluations included predominantly **large languages** with substantial digital footprint
 - These tend to be predominantly Indo-European (IE)
- We know that transfer, usually from English (which is IE) works best when transferring to other IE languages
- Datasets/benchmarks that include predominantly IE and/or large langs **overestimate** the general/global multilingual abilities of models

Multilingual Evaluation



Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). [XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning](#). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2362-2376).

- Quantifying diversity of language samples in multilingual datasets
- **Typological index**
 - Based on URIEL typological vectors of languages: 103 binary features
 - Compute entropy for each feature, and then average entropy across features
- **Family index**
 - Number of distinct language families in the language sample
- **Geography index**
 - Entropy of the distribution over the 6 global geographic macro-regions



Multilingual Benchmarks

- In the age where NLP tasks has been largely unified, it is common to evaluate models on a collection of tasks
- Multilingual benchmark: a collection of multilingual datasets
 - Not all datasets (need to) cover the same set of languages
- Some multilingual benchmarks (single task)
 - XGLUE: 11 tasks, 19 languages in total
 - XTREME: 9 tasks, 40 languages (from 12 language families)
 - XTREME-R: 10 tasks, 50 languages





Creating Multilingual Datasets

- Q: How do we normally create multilingual datasets?
 - Most commonly by translating dev/test portions of English datasets
- 1. Completely **manual** translation
 - If the original dataset has a lot of culture-specific concepts that **don't have a direct translation** or **don't exist** in the target language
 - E.g., in XCOPA: „bowling“, „parking meter“ ...
- 2. Machine translation + manual **post-editing**
 - Human annotator fixes the errors of automatic translation
 - Cheaper than manual trans. if the MT model $L_S \rightarrow L_T$ is **good enough**
- In both cases we need **bilingual annotators**
 - Difficult to find for **low-resource languages**





Model Selection in CL Transfer

- When we train ML models, we leverage a **validation** (aka **development**) dataset D_V for **model selection**
 - Selecting optimal hyperparameter values, early stopping, etc.
- When we fine-tune neural LMs for CL transfer, the **language of the validation dataset** plays a huge role
 - Target language performance much better if D_V in L_T
 - If D_V in L_S , we're selecting the model checkpoint that's optimal for the source language (usually EN) performance
- **Q:** Think of zero-shot CL transfer. What is the problem with having a validation dataset in the target language (i.e., D_V is in L_S)?





Model Selection in CL Transfer

- Q: Think of zero-shot CL transfer. What is the problem with having a validation dataset in the target language?
 - Not real zero-shot transfer! Relies on labeled instances in L_T
 - Just not directly for model training, but for model selection
- Most multilingual datasets offer both **validation** and **test** (final evaluation) data in L_T
 - Allows for unfair zero-shot transfer evaluation (labeled data in L_T)
 - Q: What if we did not need D_V in L_T for model selection?
 - We could use those $|D_V|$ in L_T for **training** instead → **few-shot** transfer
 - And few-shot is always better than zero-shot!



Model Selection in CL Transfer



Schmidt, F. D., Vulić, I., & Glavaš, G. (2023). [Free Lunch: Robust Cross-Lingual Transfer via Model Checkpoint Averaging](#). Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). *To appear*.

- **Checkpointing:** as we train the model on the training data, we **periodically** (e.g., every **N** training steps) store the parameter values
 - This is called a **checkpoint** (or **snapshot**) of the model
- **Checkpoint averaging:** the final model is the **average of all checkpoints** during training (rather than just the last checkpoint)
- Checkpoint averaging in **CL transfer** (zero-shot and few-shot) leads to more robust training behaviour and removes the need for D_V in L_T



The End

Image: Alexander Mikhailchik