

University of Ljubljana
Faculty of Computer and
Information Science



Large language models for cross-lingual transfer

Prof Dr Marko Robnik-Šikonja

15 November
2023



Contents

- Text representations
- Large language models
- Cross-lingual transfer
- Classification, summarization and QA
- Conclusions



Some images by Dan Jurafsky, Bhaskar Mitra, Nick Craswell, William Hamilton, Jacob Devlin and Jay Alammarr



Semantic language technologies

- part of everyday communication in developed countries
 - communication with mobile devices
 - intelligent search
 - digital assistants and intelligent software
 - intelligent cars and other devices
 - electronic toys
 - household appliances
 - machine translation
 - automatic summarization
 - question answering
 - writing aids
- huge progress with deep neural networks
- strong need to cover less-resourced languages



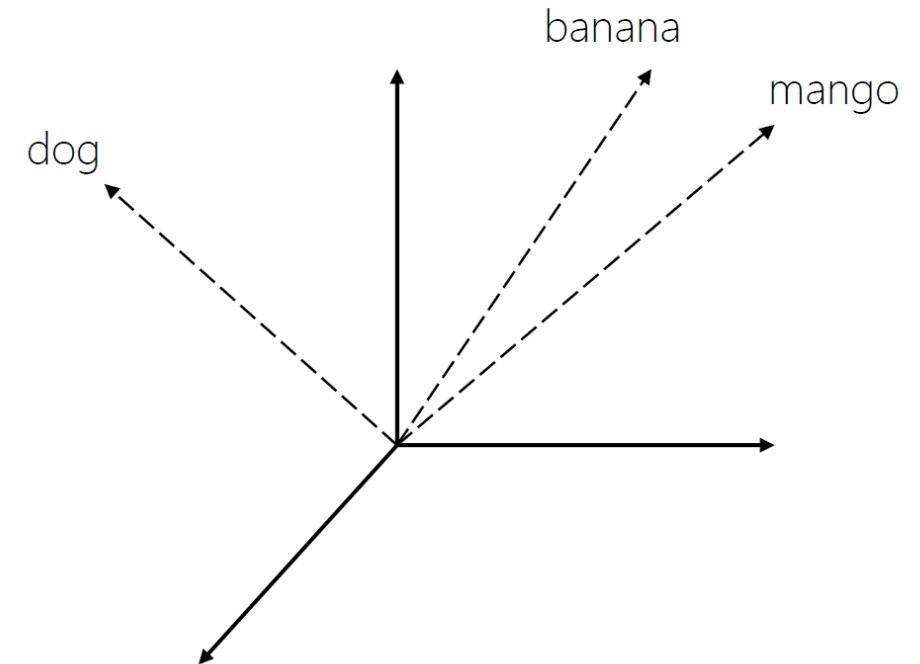
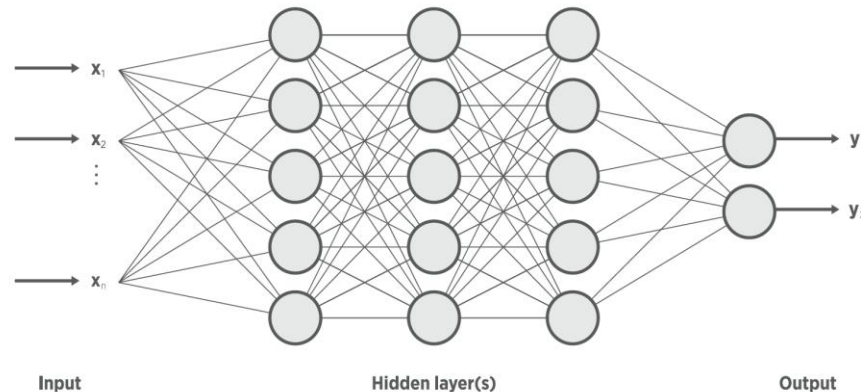
Deep neural networks for text

Currently the **most successful approach** to most natural language understanding tasks: machine translation, summarization, questions & answers, text generation, speech recognition and synthesis, etc.

Build knowledge representation automatically

Require **text as numeric representation**
the representation shall preserve
similarity and relations between words

First solution: **text embeddings** on the input



Distributional semantics



“You shall know a word
by the company it keeps”

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, p. 11. Blackwell, Oxford.

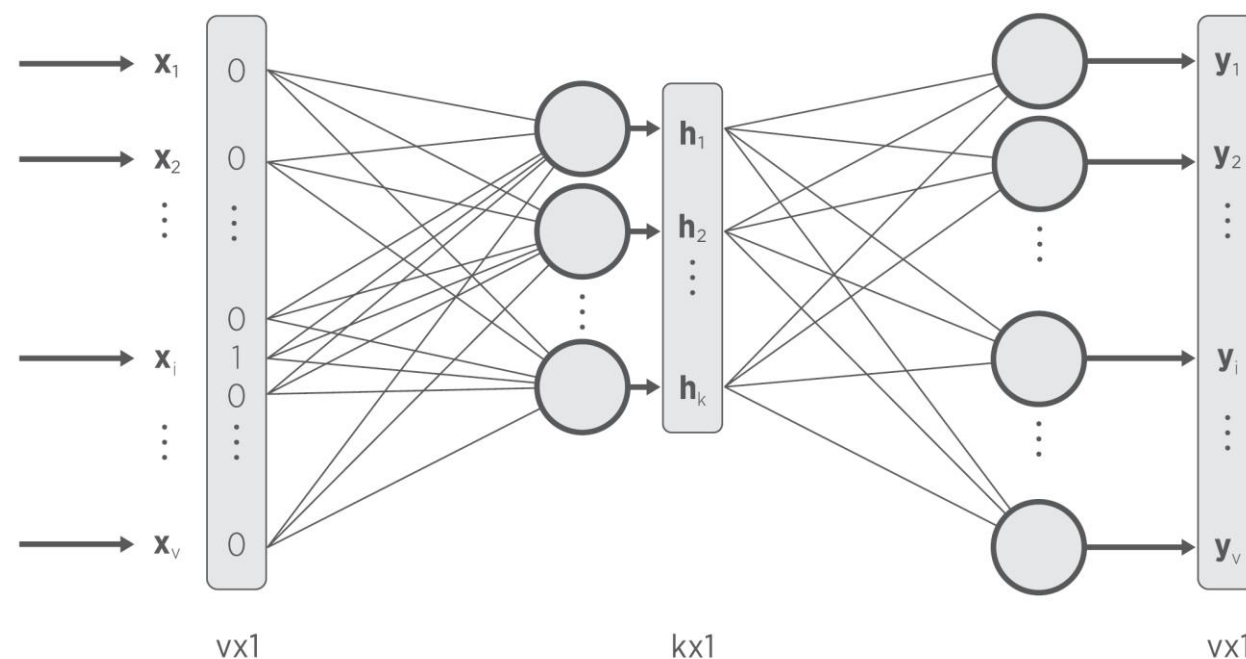
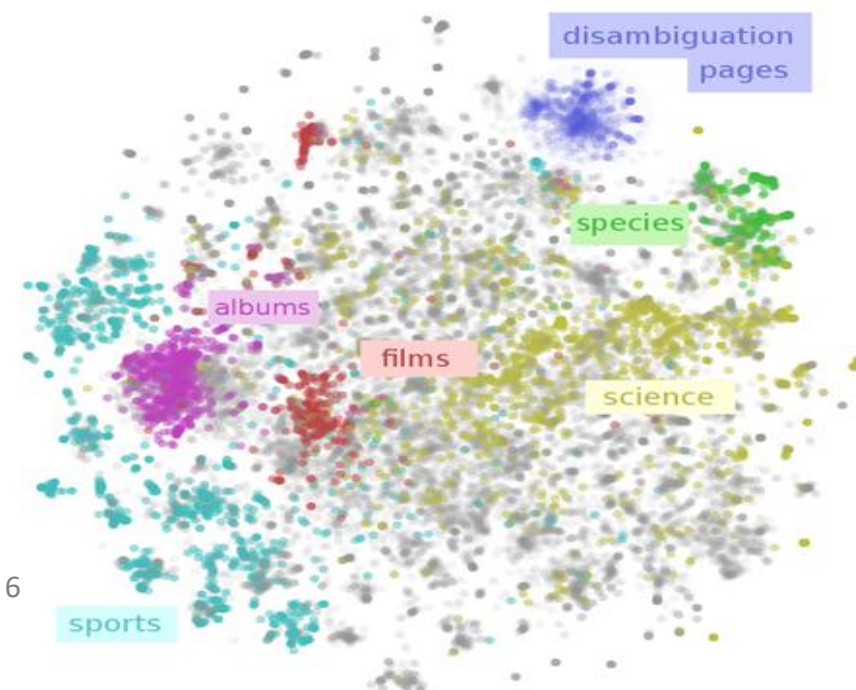


"The meaning of a word is its use in the language"

Ludwig Wittgenstein, PI #43

Word embeddings

- Representations of word meaning from **corpus statistics**
- Spatial relationships in *embedded* space correspond to **semantic relationships** expressed with language
- Called an "embedding" because it's embedded into a space
- Vectors are nowadays **learned** with neural networks



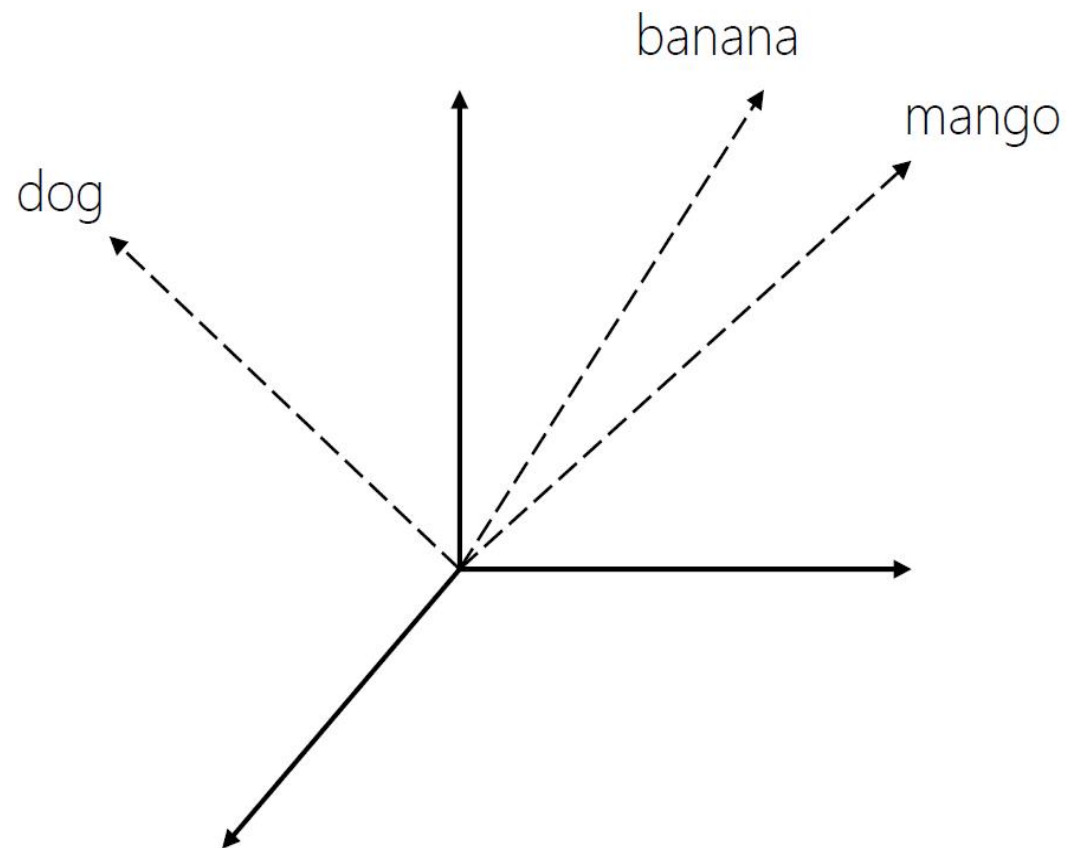
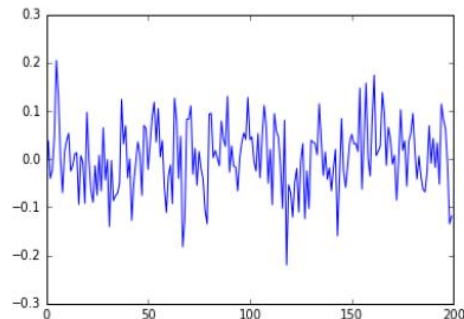
Dense embeddings

Dense. Dim = 200 (for example)

```
In [67]: print(vec['banana'])  
plt.plot(vec['banana'])
```

```
[ -0.065091,  0.037847, -0.040299, -0.022862,  0.046481,  0.204306,  0.132157,  0.000275, -0.069716,  0.014626,  0.038425,  0.053029, -  
 0.024947, -0.013991,  0.010317,  0.012735, -0.094237,  0.007101, -0.007268, -0.091869,  0.097138, -0.002357, -0.065102, -0.089856,  
 -0.013727, -0.074923,  0.007938, -0.066188,  0.064525, -0.0436, -0.001177, -0.140017, -0.003096, -0.086315, -0.0763, -0.071214,  
 -0.051458,  0.123467,  0.031151,  0.068839, -0.039029, 4e-06, -0.127185, -0.049415, -0.007708,  0.035502,  0.009538, -0.075545,  0.0  
69583,  0.062794, -0.021556,  0.031155,  0.087352,  0.117663,  0.034883,  0.104613,  0.004534,  0.037999, -0.058016, -0.110679, -0.0353  
5, -0.012488, -0.0924,  0.126315,  0.080949, -0.040334,  0.047046, -0.182169, -0.1268,  0.082376,  0.082963,  0.110073, -0.031732,  0.  
022219, -0.054332,  0.015394, -0.019853, -0.04169, -0.106969, -0.134253,  0.093094,  0.094716,  0.002643,  0.017417,  0.00309, -0.014  
145,  0.078464,  0.041464,  0.026328,  0.12988, -0.02715,  0.027002, -0.014312, -0.017305, -0.066002,  0.002747,  0.033995,  0.053829,  
  0.040628,  0.127369,  0.040216,  0.045803, -0.003395, -0.024843,  0.052411, -0.039267,  0.043378,  0.110868,  0.067947, -0.050505,  0.  
019753, -0.094825,  0.094058,  0.057547,  0.045447, -0.016258, -0.102323,  0.080506, -0.219969, -0.053595, -0.069609, -0.120579, -  
 0.048799, -0.019837, -0.109987, -0.002571,  0.031825, -0.124037, -0.024646, -0.102276,  0.038512,  0.035166,  0.031713,  0.008979,  
  0.114415,  0.0421, -0.034152,  0.014497, -0.04199, -0.018534, -0.065822, -0.020059,  0.019861, -0.159393, -0.03374,  0.083666, -0.  
025234, -0.058921, -0.014924,  0.035292,  0.050979,  0.031609,  0.0322,  0.015638,  0.146793, -0.062475,  0.042192,  0.157084,  0.00237  
1, -0.035507,  0.08275,  0.173776,  0.007175,  0.016044,  0.025942,  0.137863,  0.094541, -0.013125,  0.065621,  0.040823, -0.010574,  0.  
007796, -0.085031, -0.003617,  0.102267,  0.018047,  0.037613, -0.056187,  0.036693,  0.053867,  0.094616,  0.015941, -0.041536,  0.005  
796, -0.03694, -0.063241, -0.067796, -0.026023,  0.069142, -0.008786,  0.042428, -0.017718,  0.03318, -0.052277,  0.114012,  0.08154  
2,  0.063282, -0.012149, -0.134274, -0.118431]
```

```
Out[67]: [ <matplotlib.lines.Line2D at 0x12a60774e48>]
```



Embeddings capture relational meaning

$$\text{vector}(\text{'king'}) - \text{vector}(\text{'man'}) + \text{vector}(\text{'woman'}) \approx \text{vector}(\text{'queen'})$$



$$\text{vector}(\text{'Paris'}) - \text{vector}(\text{'France'}) + \text{vector}(\text{'Italy'}) \approx \text{vector}(\text{'Rome'})$$



Embeddings reflect cultural biases

- Ask “Paris : France :: Tokyo : x”
 - x = Japan
- Ask “father : doctor :: mother : x”
 - x = nurse
- Ask “man : computer programmer :: woman : x”
 - x = homemaker



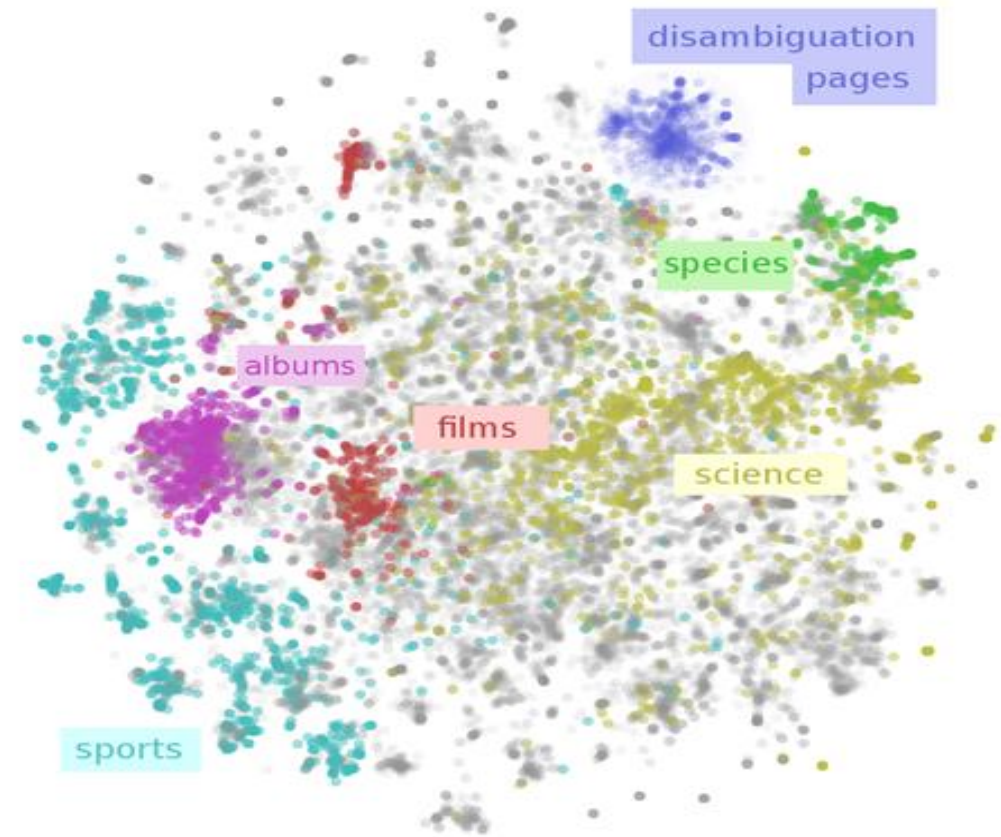
Bolukbasi, T., Kai-Wei C., Zou, J. W., Saligrama, V., and Kalai, A. W. (2016) "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357.

Ulčar, M., Supej, A., Robnik-Šikonja, M., & Pollak, S. (2021). Slovene and Croatian word embeddings in terms of gender occupational analogies. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 9(1), 26-59.

Cross-lingual embeddings

- embeddings are trained on monolingual resources
- words of one language form a cloud in high dimensional space
- clouds for different languages can be aligned

$$W_1 S \approx E$$



Cross-lingual embeddings

- Aligning embedding spaces across languages

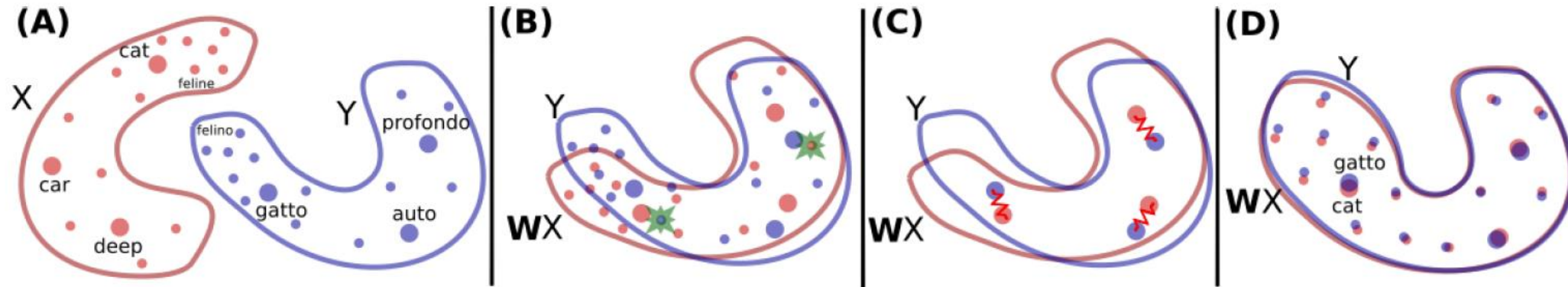


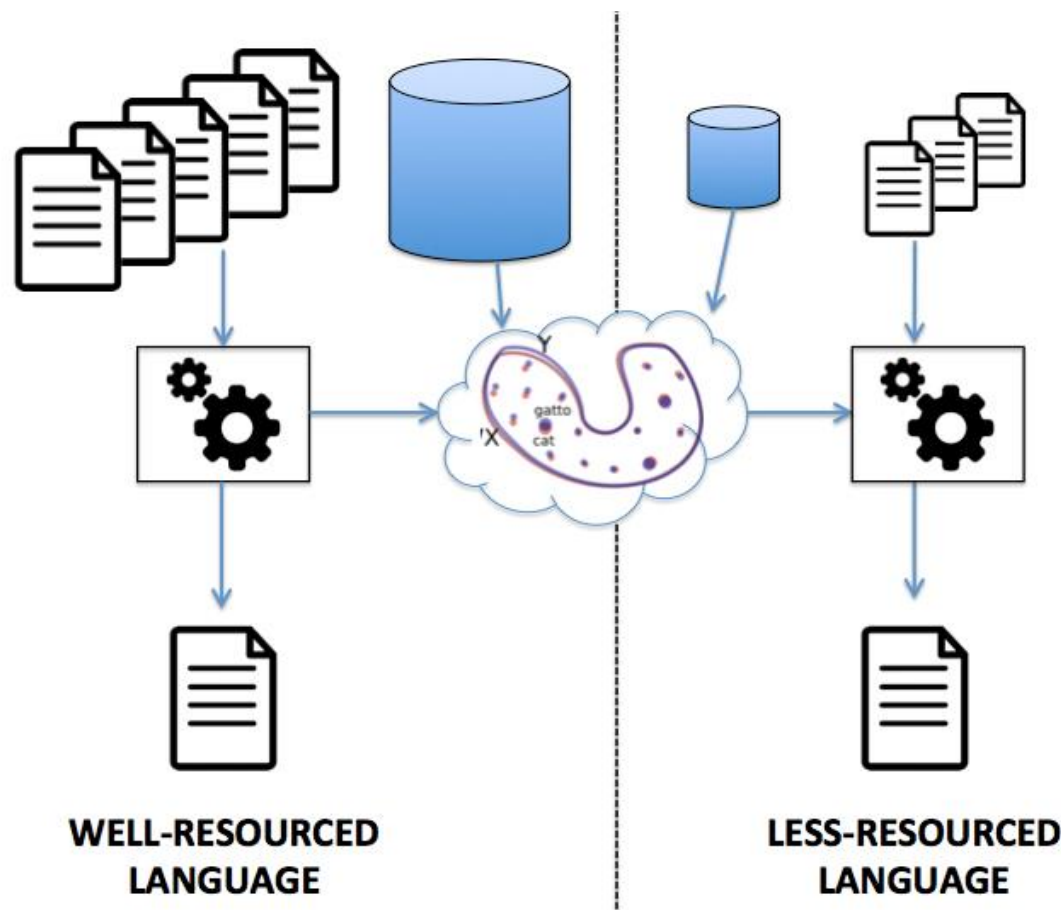
Image credit to Conneau, Lample, Ranzato, Denoyer, Jegou: Word translation without parallel data. ICLR 2018.

- Supervised, semi-supervised, and unsupervised approaches

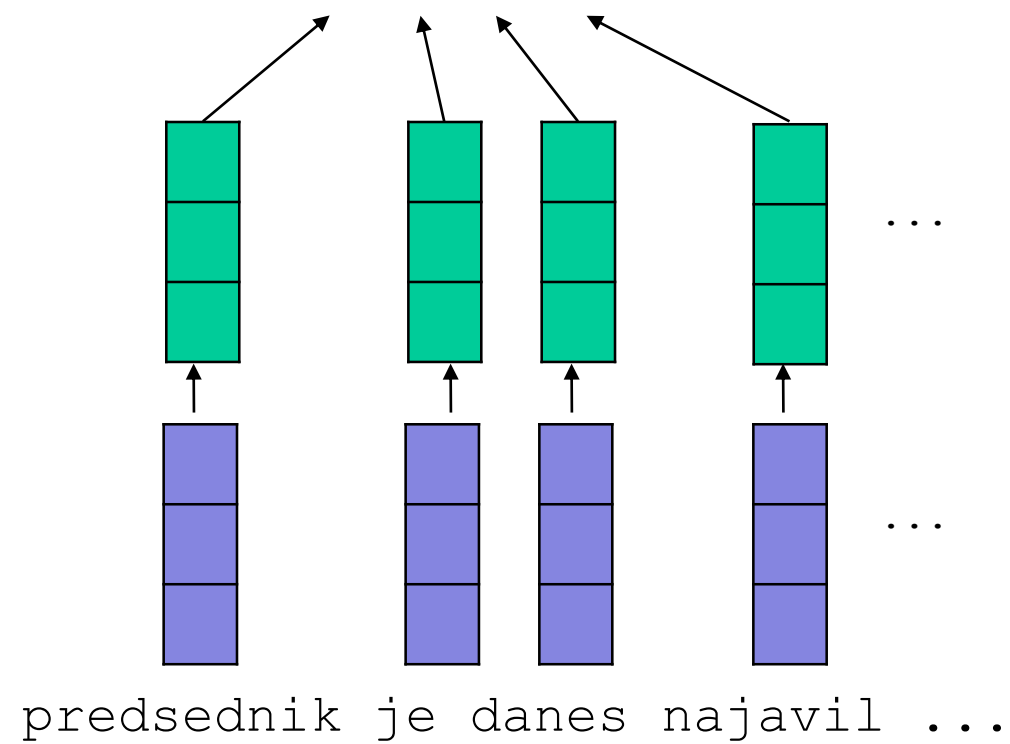
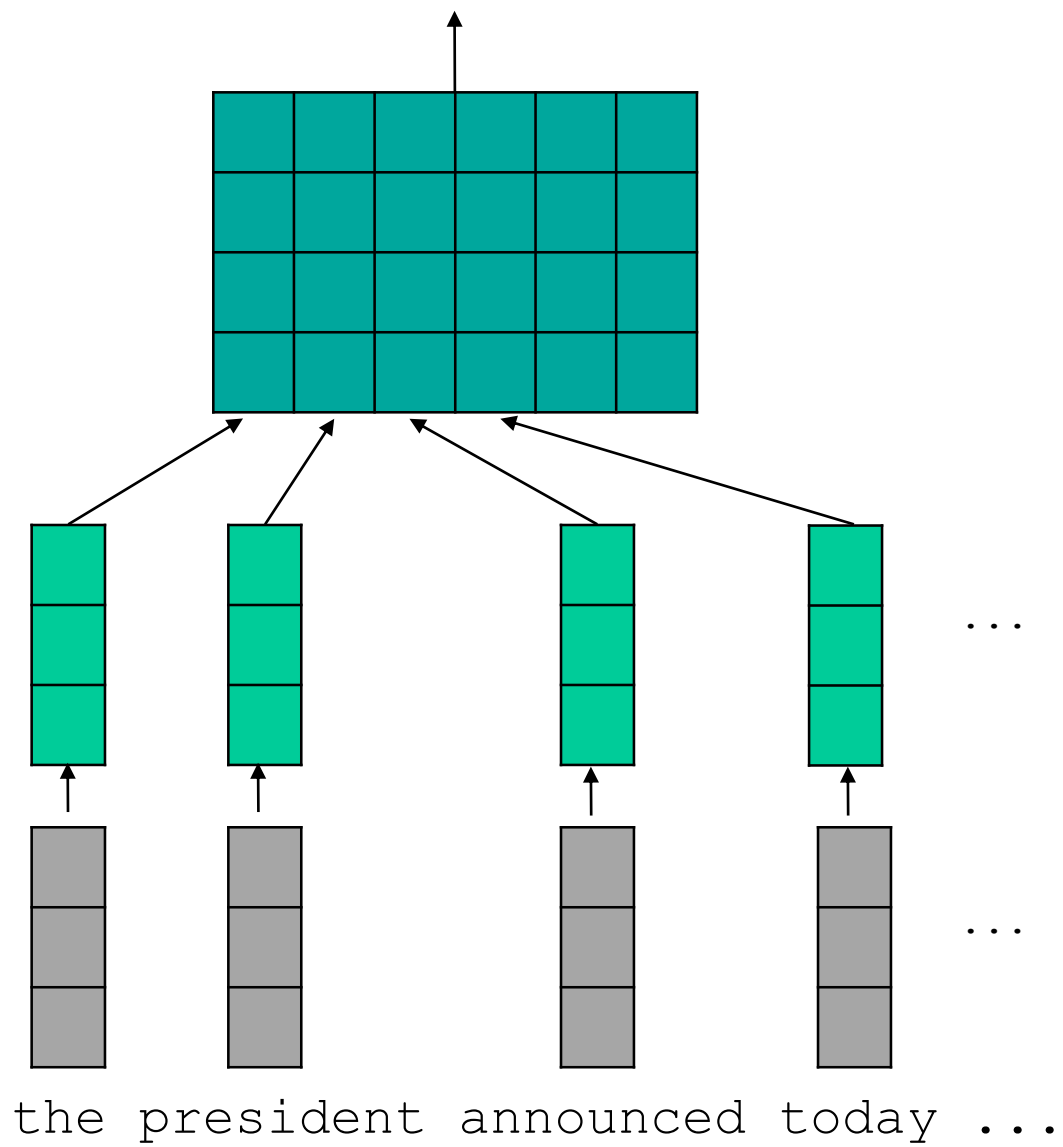
Ruder, S., Vulić, I. and Søgaard, A., 2019. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, 65, pp.569-631.

Cross-lingual model transfer based on embeddings

- Transfer of tools trained on mono-lingual resources



Often works better than machine translation



Analogies in many languages and cross-lingually

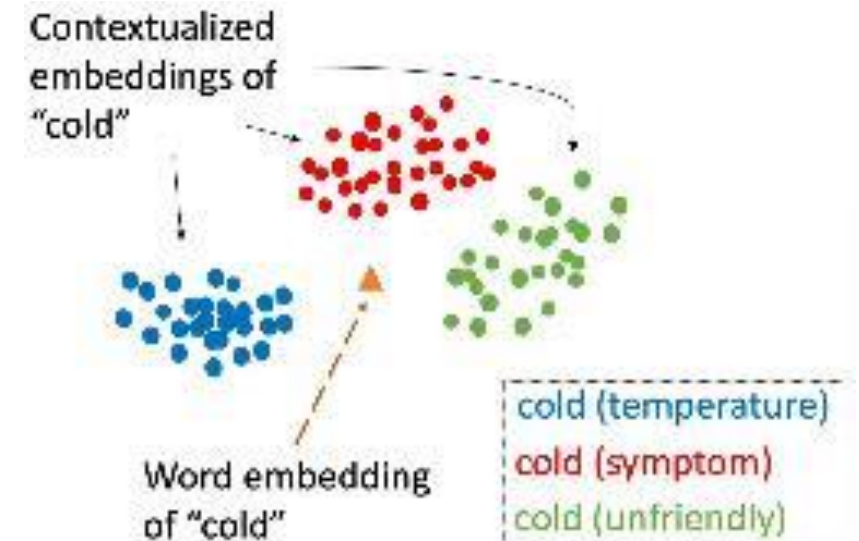
- First analogy dataset constructed by Mikolov et al. (2013)
- 15 categories: capitals, family, city with river, animal genus, comparative adjective, superlative adjective, adjective-adverb, etc
- Culturally adapted to many languages, e.g., Croatian, English, Estonian, Finnish, Latvian, Lithuanian, Russian, Slovenian, and Swedish
- Cross-lingual test “father : doctor :: mama : x”
- x = medicinska sestra

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint 1301.3781.

Ulčar, Matej, Kristiina Vaik, Jessica Lindström, Milda Dailidenaite, and Marko Robnik-Šikonja. “Multilingual Culture-Independent Word Analogy Datasets”. In: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4067–4073.

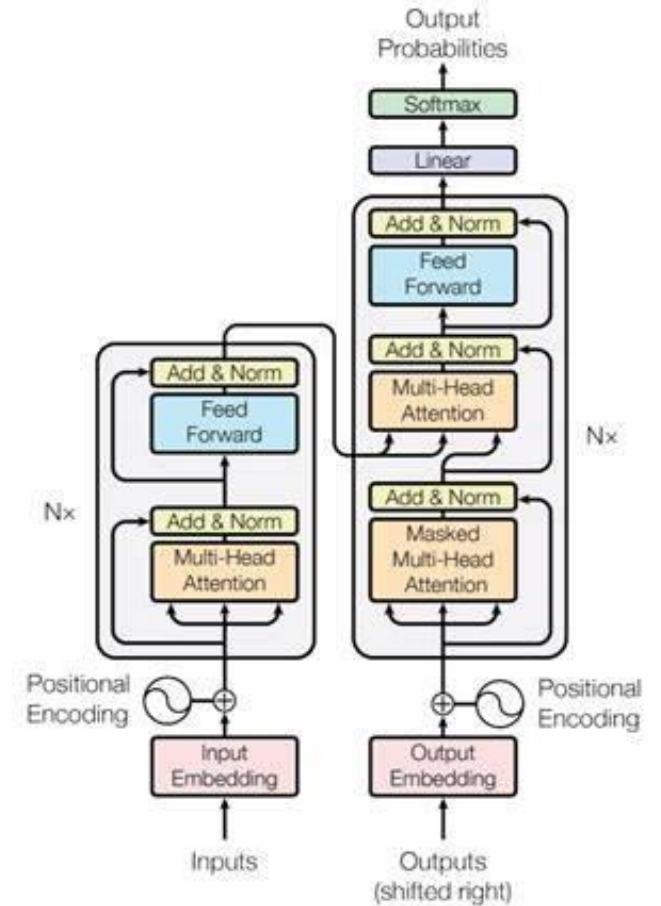
Contextual embeddings

- **Static embeddings**, like word2vec, produce the **same vector** for a word like “cold” irrespective of its meaning and context
- Contextual embeddings like ELMo and BERT take the **context** into account



LLMs for text processing

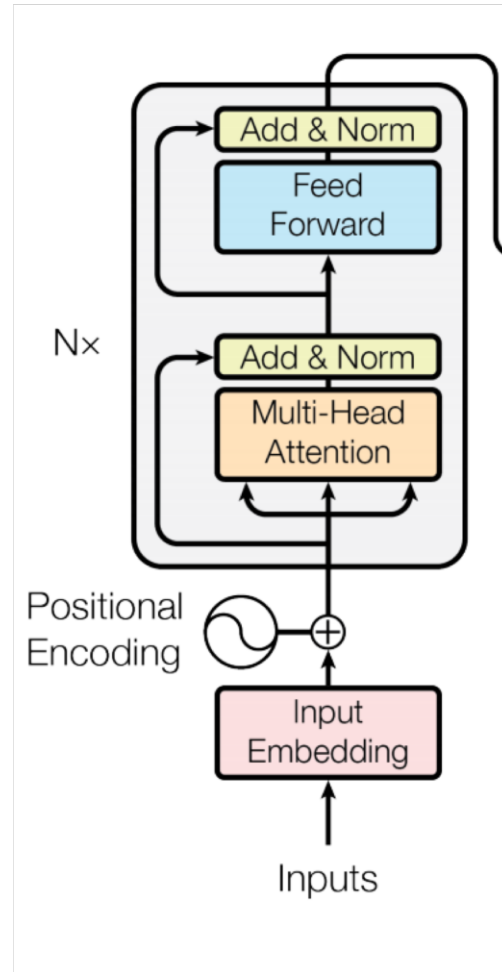
- LLMs based on transformer neural networks
- Learning in two phases
 - The first phase: pretraining
 - predict masked word, next word, etc.
 - huge amounts of text
 - long training
 - The second phase: fine-tuning for specific task
 - much faster than the first phase
 - transfer of knowledge about language from the first phase



BERT & co.

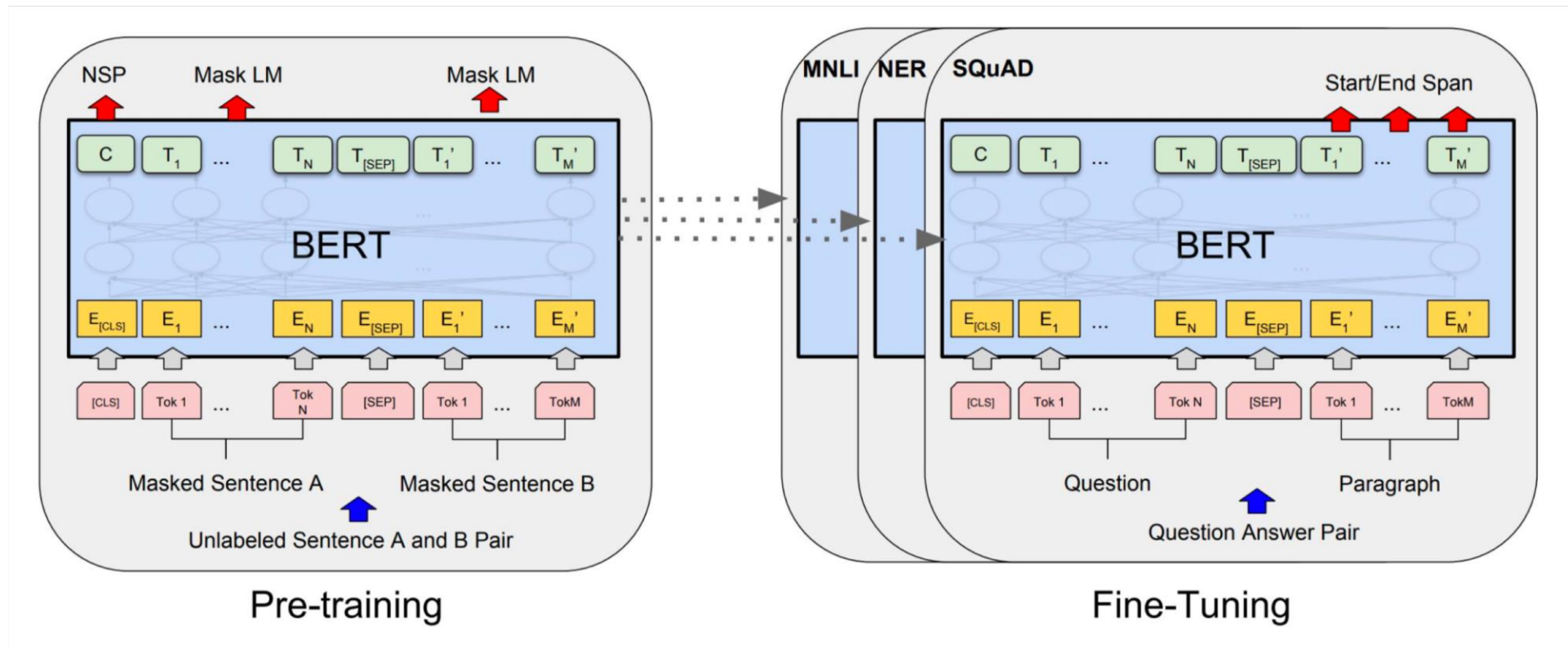
- BERT is pretrained as masked language model
- Uses the encoder part of the transformer neural architecture
- Deep with e.g., 12/24 layers, 110/340 million weights
- Originally available for English and Chinese
- Computationally intensive pretraining
- Several variants: RoBERTa, ALBERT, DistilBERT, Electra, etc.
- Hundreds of papers investigating BERT-like models

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.



Use of BERT

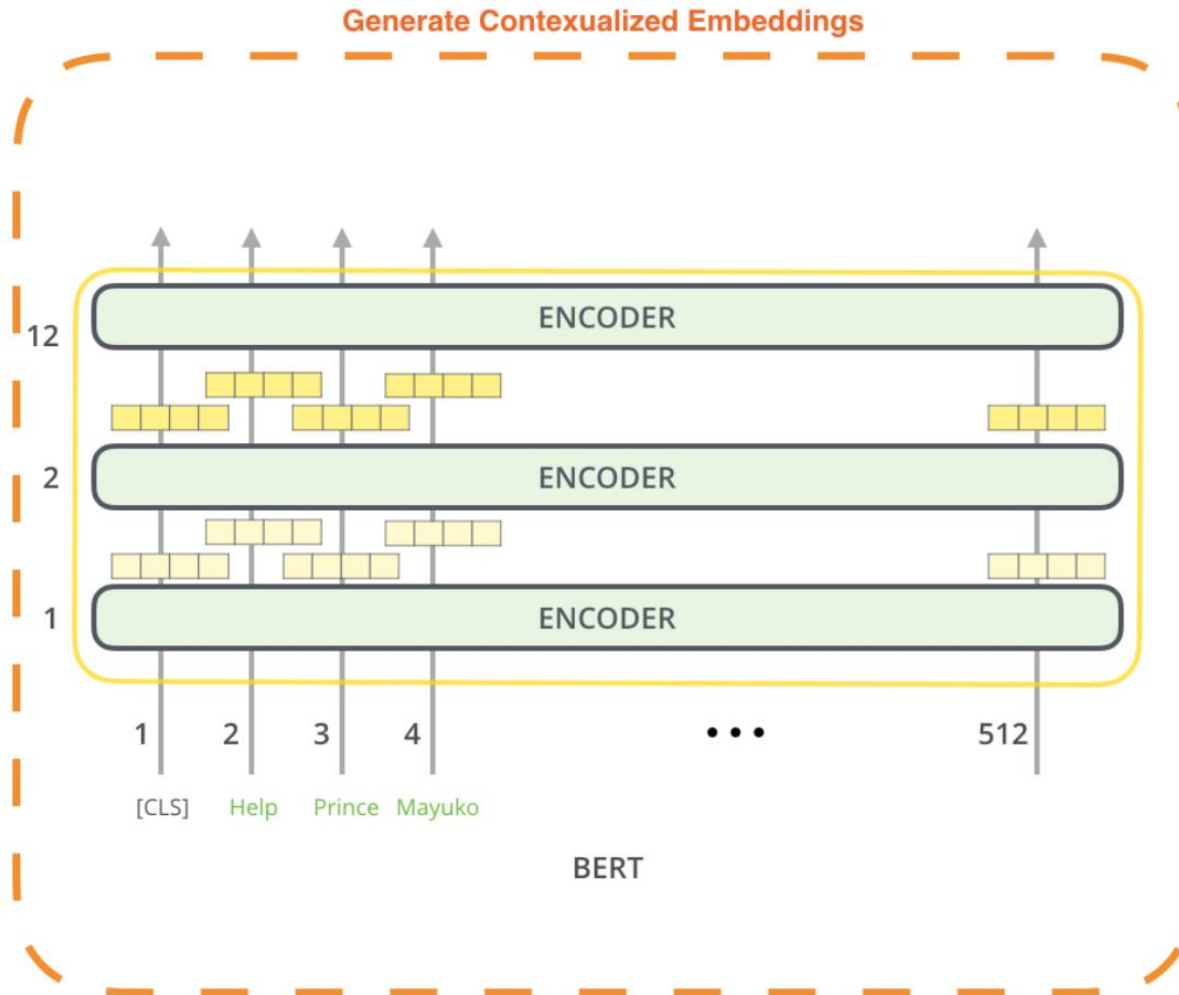
- train a classifier built on the top layer for each task that you fine tune for, e.g., Q&A, NER, inference
- achieves good results for many tasks
- GLUE and SuperGLUE tasks for NLI



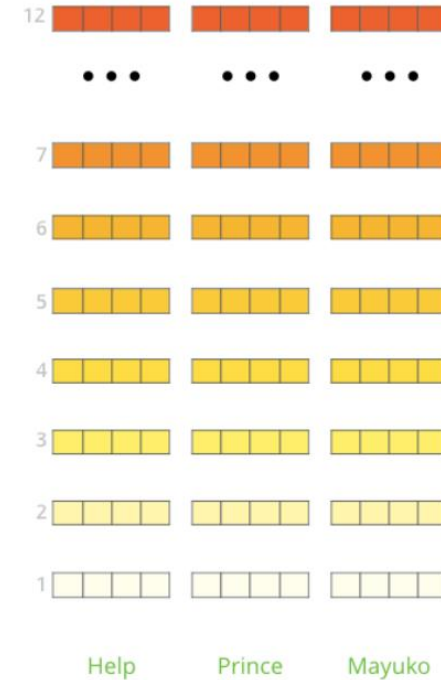
BERT can produce embeddings

- one can extract fixed size contextual vectors from BERT, achieving slightly lower accuracy than using the whole BERT as the first stage model

Layer-wise embeddings












The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

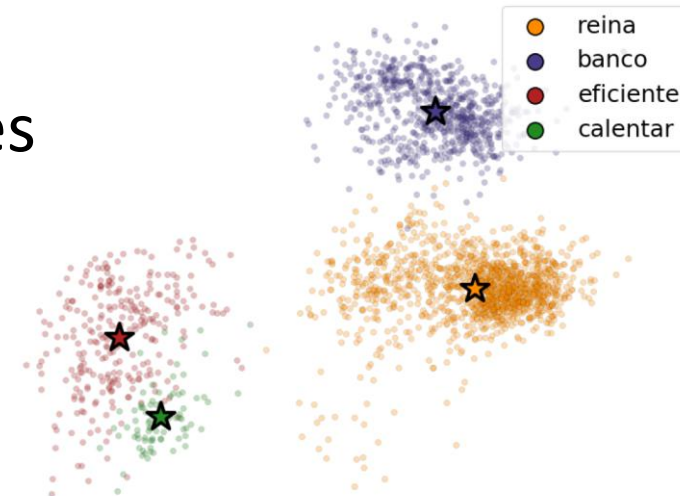
Which layer of BERT to use as embeddings?

What is the best contextualized embedding for “**Help**” in that context?
 For named-entity recognition task CoNLL-2003 NER

		Dev F1 Score
12 	First Layer	91.0
• • •		
7 	Last Hidden Layer	94.9
6 		
5 	Sum All 12 Layers	95.5
4 		
3 	Second-to-Last Hidden Layer	95.6
2 		
1 	Sum Last Four Hidden	95.9
		
Help	Concat Last Four Hidden	96.1

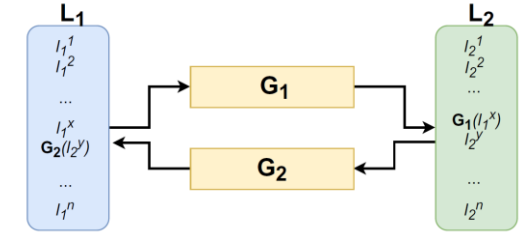
Cross-lingual contextual embeddings

- Problem: contextual embeddings produce a different vector for each word in context (e.g., a sentence)
- Bilingual and multilingual dictionaries can provide anchoring points for alignment of different words but without contexts
- Alignment of contextual embeddings
 - bilingual and multilingual corpora of parallel sentences
 - sentences have to be aligned
 - alignment of word clusters
 - non-isomorphic alignment

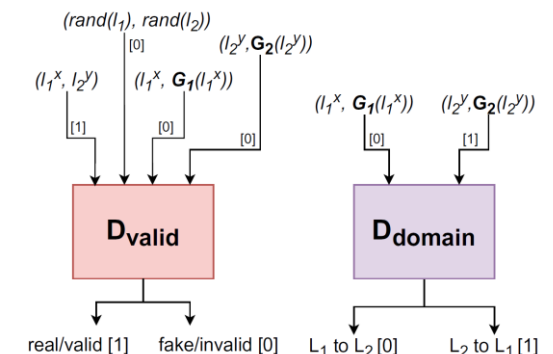


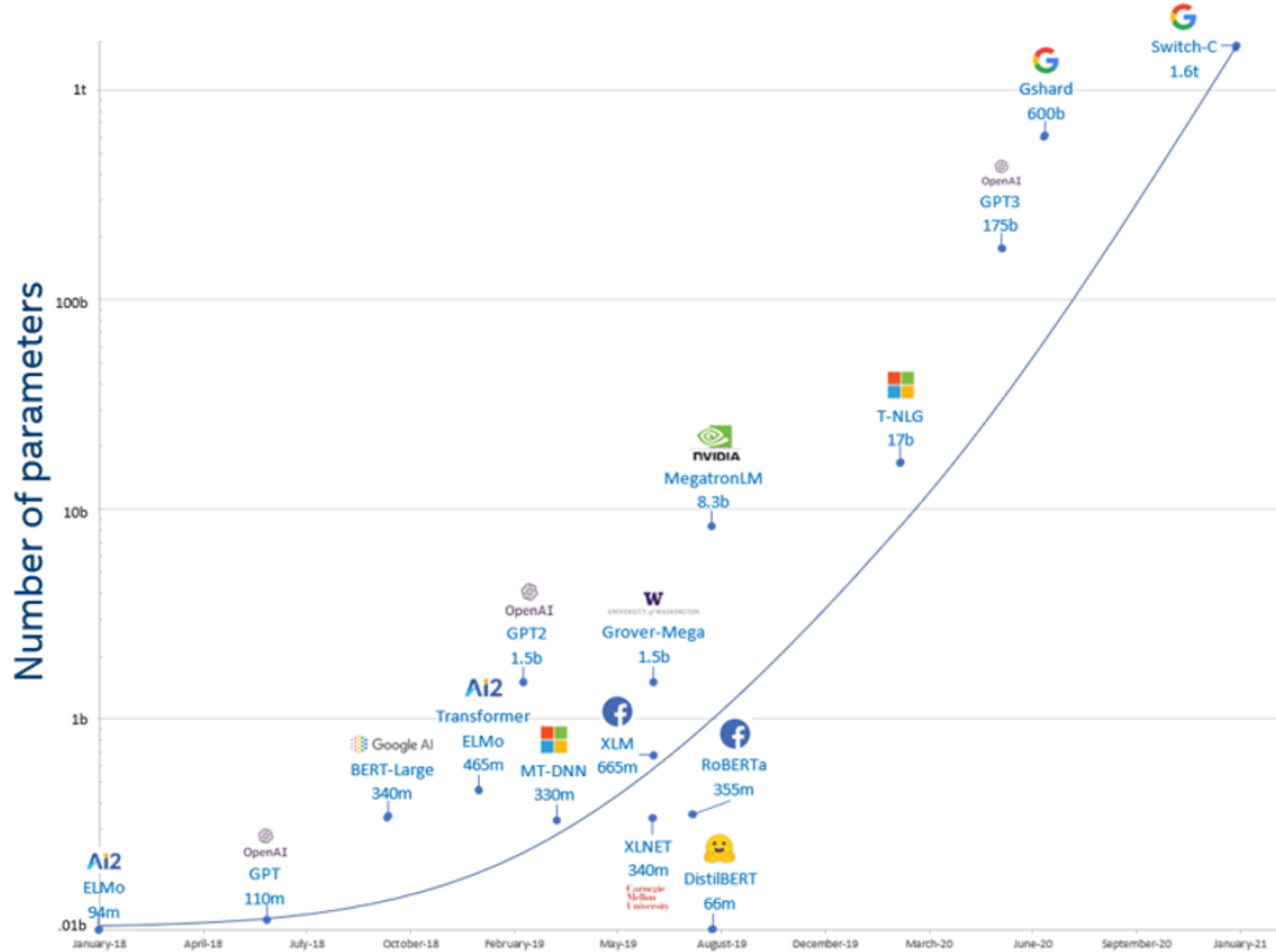
Non-isomorphic cross-lingual embeddings

- Cross-lingual maps for contextual embeddings
 - Parallel sentences based alignment dataset
 - Many different types of anchor points between languages: low- and high-quality dictionaries, named entities, and linked entities obtained from BabelNet
 - Isomorphic maps based on Vecmap and MUSE libraries
 - Non-isomorphic maps using GANs
- Results for ELMo embeddings over several tasks:
 - Isomorphic Vecmap and non-isomorphic ELMoGAN methods work best
 - The best method is dataset dependent, requires fine-tuning of the methods' hyperparameters



Ulčar, M. and Robnik-Šikonja, M. (2022). "Cross-lingual alignments of ELMo contextual embeddings". *Neural Computing and Applications* DOI: <https://doi.org/10.1007/s00521-022-07164-x>

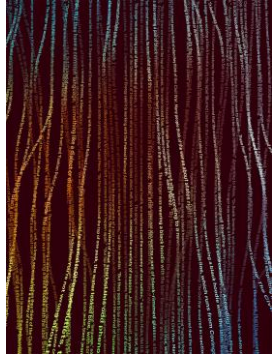






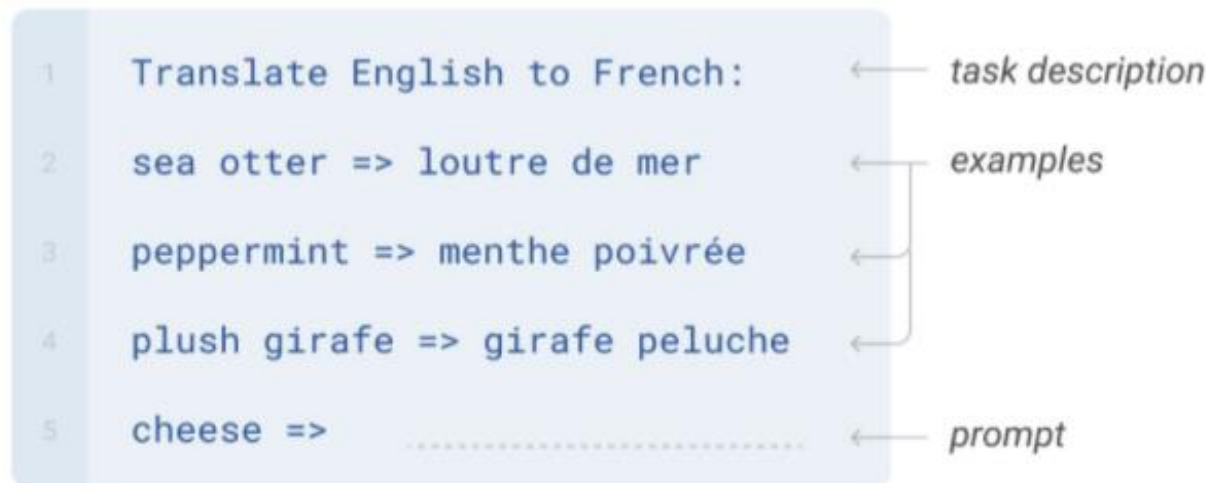
Huge generative language models

- ChatGPT, OpenAI, Nov. 2022
based on GPT-3.5 with additional training for dialogue
- uses RLHF (reinforcement learning with human feedback)
- demo: <https://chat.openai.com/>
- huge public impact, possibly disruptive for writing professions, learning, teaching, scientific writing
- GPT-4, 2023: even larger, allows longer context, image input
- excellent on generative tasks, lag in classification tasks
- cross-lingual to some degree
- LLaMa-2, Alpaca, Koala, LiMa, Falcon: smaller, very competitive open models, support very few languages



In-context learning in LLMs

- use text input to condition the model on task description and some examples with ground truth.
- Uses zero-shot learning, one-shot learning, few-shot learning (examples have to fit into the context window, usually 10-100)
- no gradient updates are performed
- can be used for cross-lingual transfer in multilingual models

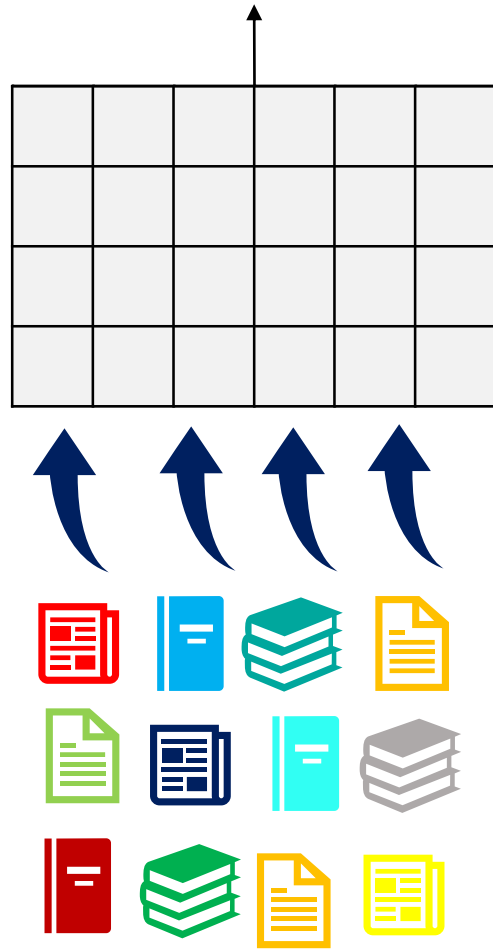




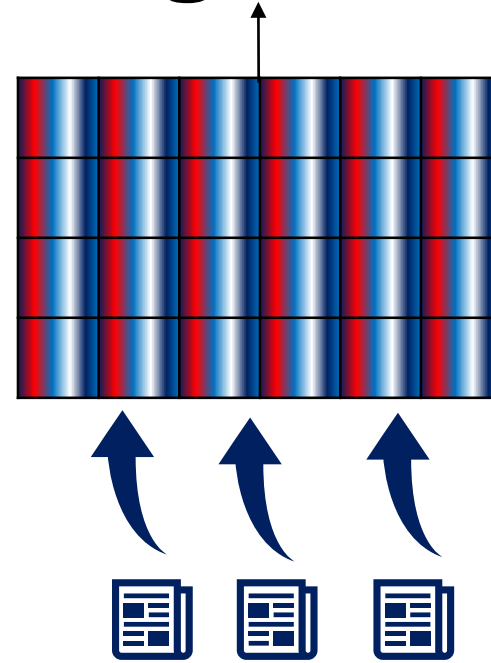
Multilingual PLMs

- Pretrained on multiple languages simultaneously
- multilingual BERT supports 104 languages by training on Wikipedia
- XLM-R was trained on 2.5 TB of texts
- allow cross-lingual transfer
- solve problem of insufficient training resources for less-resourced languages

Using multilingual models



Pretraining



Fine-tuning

predsednik je danes najavil ...

Classification



Zero-shot transfer and few-shot transfer

Why not only multilingual?

- performance on many tasks drops with more languages
- results for a few tasks in Slovene (Named Entity Recognition – NER, Part-of-Speech Tagging – POS, Dependency Parsing – DP, Sentiment Analysis – SA, Word Analogy – WA)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
SloBERTa	0.933	0.991	0.844	0.623	0.405



Dictionaries in PLMs

- Tokenization depends on the dictionary
- The dictionary is constructed statistically (e.g., Byte-Pair Encoding or SentencePiece algorithm)
- Sentence: “Letenje je bilo predmet precej starodavnih zgodb.”
- SloBERTa:
'_Le', 'tenje', '_je', '_bilo', '_predmet', '_precej', '_staroda', 'vnih', '_zgodb', '.'
- mBERT:
'Let', '##en', '##je', 'je', 'bilo', 'pred', '##met', 'pre', '##cej', 'star', '##oda', '##vnih', 'z', '##go', '##d', '##b', '.'



Trade off: fewlingual models

- BERT trained with only a few languages
- more data for training
- more specific dictionary
- good for cross-lingual transfer
- Trilingual models
 - CroSloEngual BERT
 - FinEst BERT
 - LitLat BERT
- Bilingual and fewlingual
 - SIEng BERT
 - SlavBERT (ru, pl, cs, bg; DeepPavlov)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
CSE-BERT	0.928	0.990	0.854	0.610	0.195
SloBERTa	0.933	0.991	0.844	0.623	0.405

- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In International Conference on Text, Speech, and Dialogue (pp. 104-111).

XL transfer in classification

- Excellent XL transfer between similar languages like Slovene and Croatian

sentiment analysis

Source	Target	LASER		mBERT		CSE BERT		Both target	
		\bar{F}_1	CA	\bar{F}_1	CA	\bar{F}_1	CA	\bar{F}_1	CA
Croatian	Slovene	0.53	0.53	0.53	0.54	0.61	0.60	0.60	0.60
Croatian	English	0.63	0.63	0.63	0.66	0.62	0.64	0.62	0.65
English	Slovene	0.54	0.57	0.50	0.53	0.59	0.57	0.60	0.60
English	Croatian	0.62	0.67	0.67	0.63	0.73	0.67	0.73	0.68
Slovene	English	0.63	0.64	0.65	0.67	0.63	0.64	0.62	0.65
Slovene	Croatian	0.70	0.65	0.64	0.63	0.73	0.69	0.73	0.68
Croatian English	Slovene	0.54	0.54	0.53	0.54	0.60	0.58	0.60	0.60
Croatian Slovene	English	0.62	0.61	0.65	0.67	0.63	0.65	0.62	0.65
English Slovene	Croatian	0.64	0.68	0.63	0.63	0.68	0.70	0.73	0.68
Average performance gap		0.04	0.03	0.04	0.03	0.00	0.01		

idiom detection

Language	Slovene ELMo	mBERT	Default F_1
Slovene	0.8163	0.8359	0.667
Croatian	0.9191	0.8970	0.667
Polish	0.2863	0.6987	0.667

- Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 9(1), 1-25.
- Škvorc, T., Gantar, P., & Robnik-Šikonja, M. (2022). MICE: mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235, 107606.



What XL LLMs learn?

- We would like to travel to [MASK], ki je najlepši otok v Mediteranu.

SloBERTa: ..., Slovenija, I, Koper, Slovenia

CSE-BERT: Hvar, Rab, Cres, Malta, Brač

XLM-R: Mallorca, Tenerife, otok, Ibiza, Zadar

mBERT: Ibiza, Gibraltar, Tenerife, Mediterranean, Madeira

BERT (en): Belgrade, Italy, Serbia, Prague, Sarajevo





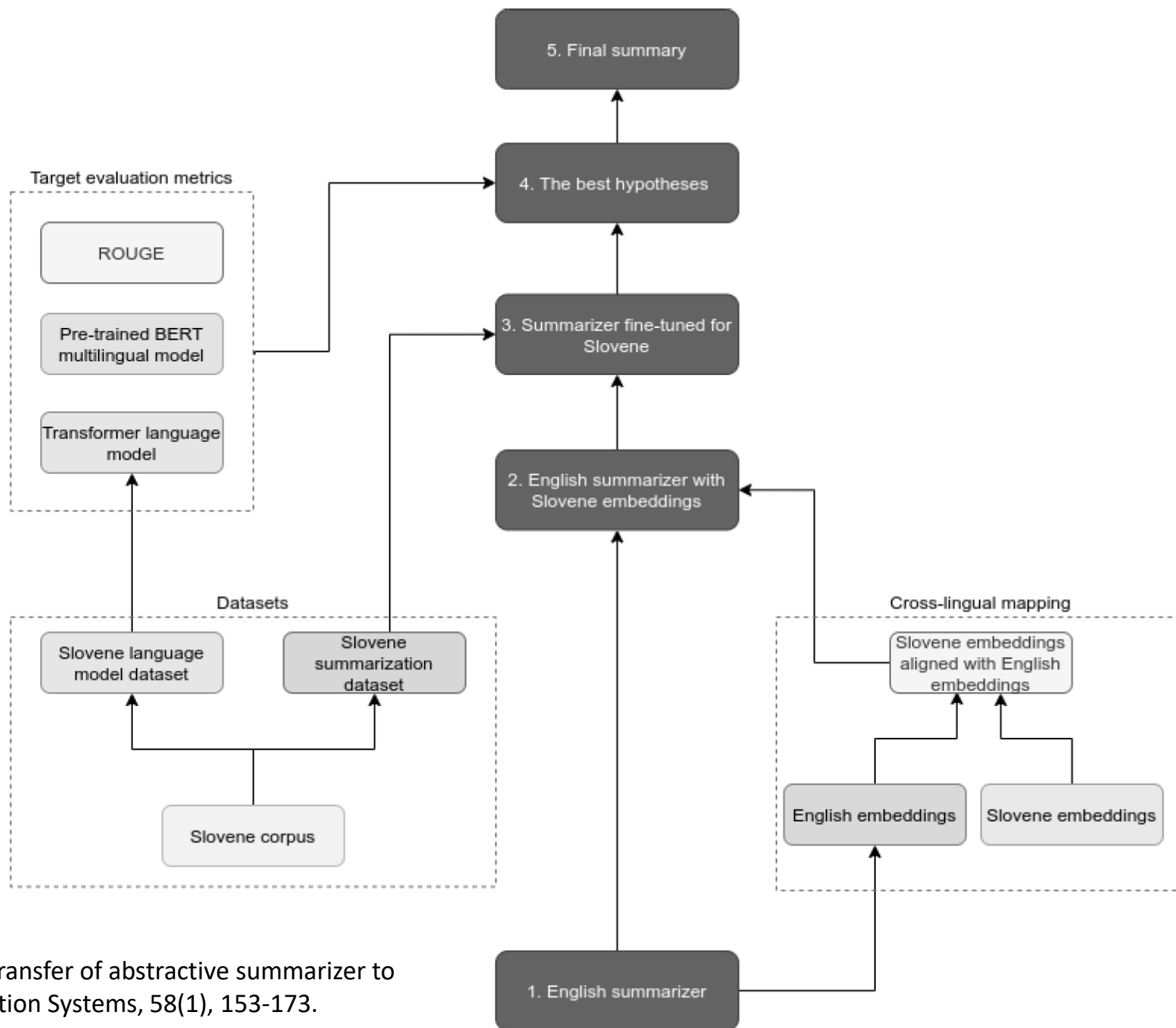
XL summarization

- just using cross-lingual transfer won't work well (why?)
- but still: can we use a trained English model in another language?
- yes, if we polish the output
- polishing: language model in another language, fine-tuning
- we used the English model by Chen and Bansal (2018), which is an encoder-decoder architecture, combining an extractive and abstractive summarizer, with RL
- evaluation on Slovene news
 - zero shot transfer
 - a few shot transfer

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, pages 675–686



XL summarization architecture





XL summarization: evaluation setting

- evaluation on Slovene news
 - zero shot transfer
 - a few shot transfer
 - no transfer
- 287,226 English summaries, 117,563 Slovene summaries

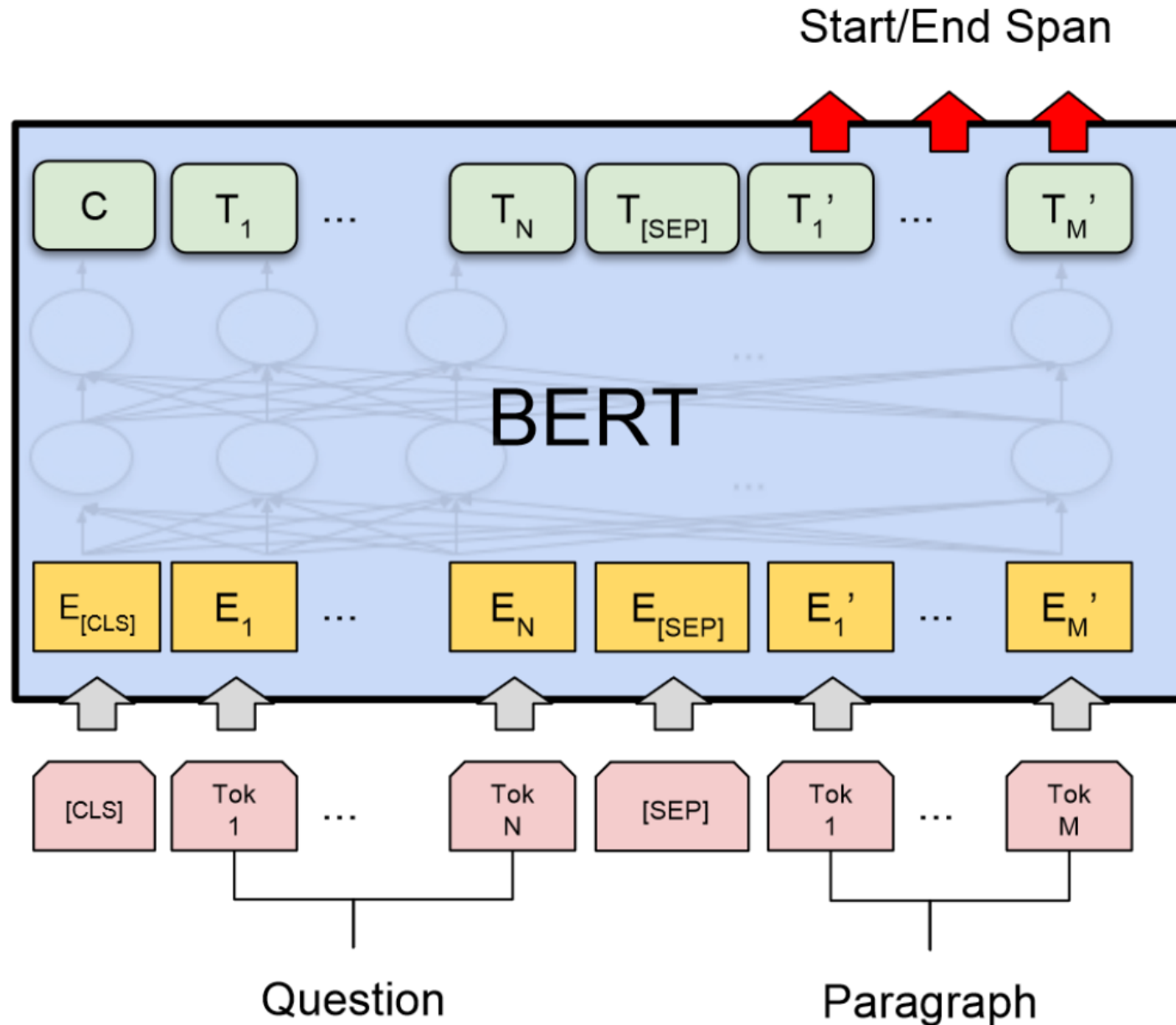
Model	Slovene data size		Details
	in %	# instances	
MENG	0%	0	cross-lingual mappings, no fine-tuning, zero-shot transfer
M1	1%	1,176	cross-lingual mappings, trained extractor, fine-tuned abstractor
M10	10%	11,756	cross-lingual mappings, trained extractor, fine-tuned abstractor
M25	25%	29,391	cross-lingual mappings, trained extractor, fine-tuned abstractor
M50	50%	58,782	cross-lingual mappings, trained extractor, fine-tuned abstractor
M100	100%	117,563	cross-lingual mappings, trained extractor, fine-tuned abstractor
MSLO	100%	117,563	Slovene embeddings, trained extractor, trained abstractor, no transfer



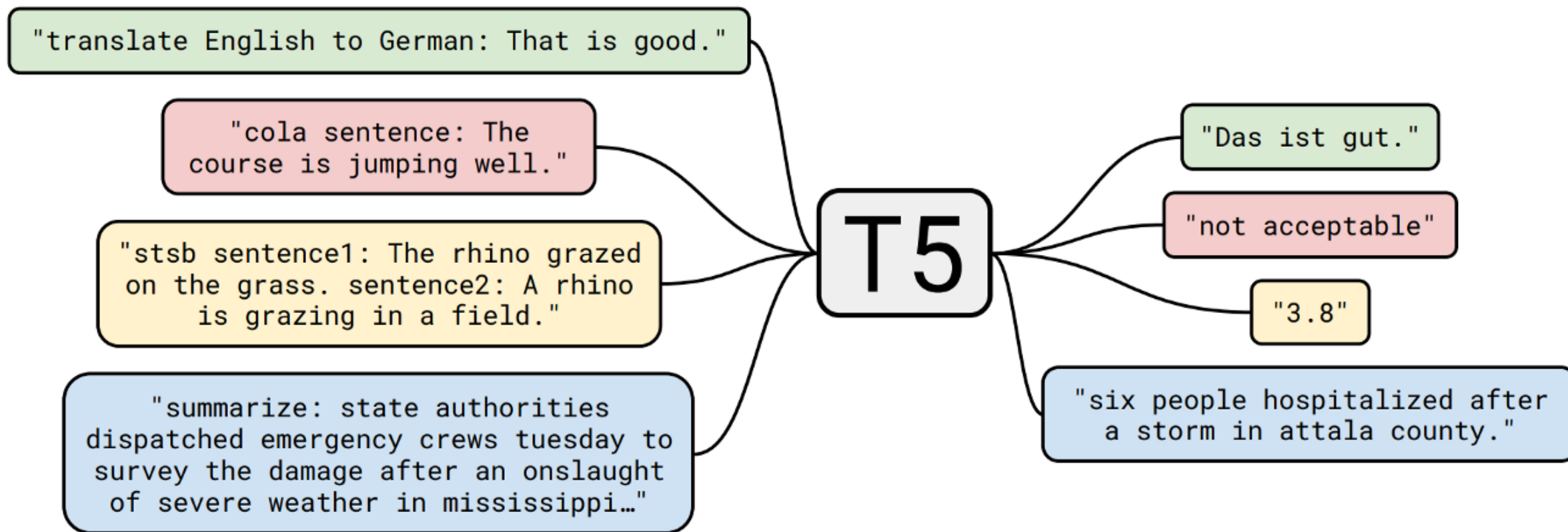
XL summarization: results

Model	Average generated		Evaluation scores			
	sentences	characters	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity
MENG	3,99	500,61	18,91	3,74	16,27	3,69
M1	2,81	218,48	12,94	1,96	11,61	4,23
M10	1,95	204,59	15,71	3,71	13,87	2,14
M25	2,89	159,00	19,32	5,00	17,12	2,19
M50	3,01	168,59	21,30	6,09	18,91	2,15
M100	2,79	297,67	21,67	6,81	19,16	2,12
MSLO	2,58	270,79	21,07	6,62	18,64	2,13
Reference Slovene	2,10	302,02				
Reference English	3,88	312,51				
ROUGE-L & BERTScore			24,97	7,43	21,50	

Questions and answers with BERT



T5 (Text-To-Text Transfer Transformer) models



Unified QA

- Use several types of questions in T5 model to generate answers: extractive, abstractive, multichoice, yes/no
- A model is trained on all types of questions
- Fine-tuned on a specific type of questions

Extractive [SQuAD]

Question: At what speed did the turbine operate?

Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?

Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...

Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?

Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar

Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?

Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...

Gold answer: no

Khashabi, Min, Khot, Sabharwal, Tafjord, Clark in

Hajishirzi. UnifiedQA: Crossing Format Boundaries With a Single QA System. V *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, str. 1896–1907, 2020.

Unified QA in Slovene

- Use partially human translations of English QA datasets to Slovene (mostly taken from Slovene SuperGLUE benchmark)
- use SloT5 model and mT5 model
- quantitatively slightly worse than English model
- qualitative analysis:
 - the generated answers are mostly substrings or given choices in multiple-choice questions
 - models cannot paraphrase, rephrase or provide answers in the correct Slovene case
 - problems with multi-part questions requiring multiple answers that are not listed in the same place in the context
 - machine translations are not always grammatically correct or do not make it clear what the question is asking for
 - best performance on factoid questions that require a short answer

- Ulčar, M., and Robnik-Šikonja, M. (2022) Sequence to sequence pretraining for a less-resourced Slovenian language. *arXiv preprint arXiv:2207.13988*, 2022.
- Žagar, A., & Robnik-Šikonja, M. (2022). Slove ne SuperGLUE Benchmark: Translation and Evaluation. Proceedings of LREC 2022.
- Logar, K. and Robnik-Šikonja, M. (2022) Unified Question Answering in Slovene. Proceedings of IS 2022: Slovene Artificial Intelligence Conference, SCAI 2022.



Conclusions

- Cross-lingual transfer has huge potential for less-resourced languages
- Current technology covers a few hundred most resourced languages
- What about the others?
- Attempts to make LLMs less resource hungry.
- Lots of XL applications
- Lots of opportunity for collaboration between NLP researchers in similar languages

