Transformers and multilinguality

Содержание курса

1. Архитектура Transformer
2. Предобученные трансформеры: BERT и его друзья
3. Решение задач NLU с помощью трансформеров
4. Трансформеры для генерации текста
5. Решение sequence-to-sequence задач с помощью трансформеров
6. Мультиязычные трансформеры
7. Сжатие и ускорение трансформеров
8. Трансформеры для работы с графами
9. Мультимодальные и картиночные трансформеры
10. Трансформеры для последовательностей событий

Agenda

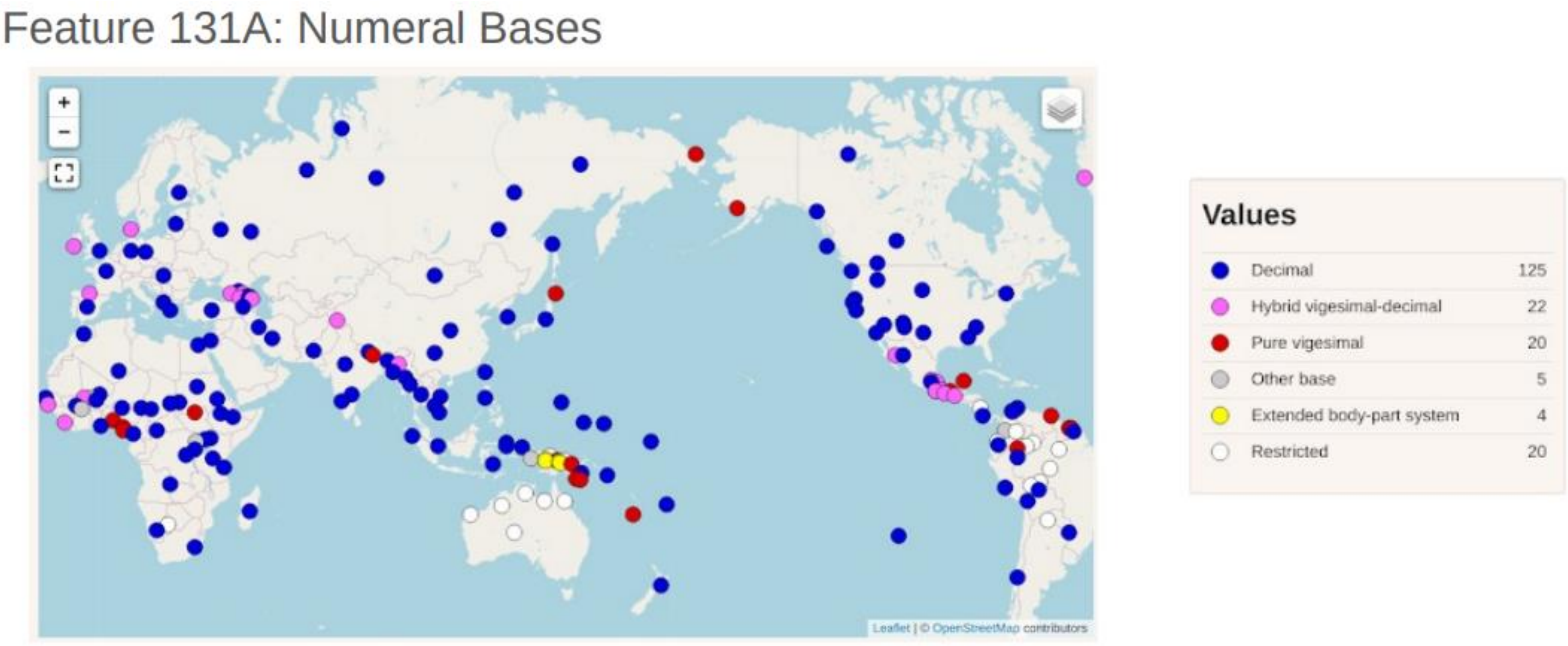
- Multilingual resources, tasks and difficulties
- Sentence encoders (from lecture 3)
- Text retrieval (from lecture 3)
- Multilingual models
- Tricks and recipes for multilingual NLP

Why multilingual?

- Languages other than English or Russian do exist.
  - And as empires fall apart, new languages get official status and wider usage
  - Users want their content in their own languages
- It is expensive to support separate NLP models for each language
- Most languages are “low-resource”
  - Monolingual models for them are often not good enough
  - But we can transfer NLP knowledge across languages
    - For closely related languages (e.g. ru->by), it can be transferred directly
    - For more distant languages, translation might be required
      - For good translation, we need parallel corpora to train
      - To collect parallel corpora, we need good NLU models…

The language space: WALS typology

The World Atlas of Language Structures stores unified language *features*



http://phontron.com/class/multiling2022/assets/slides/multiling-03-typology.pdf ; https://wals.info/

The language space: WALS typology

- 2,676 languages, 192 attributes

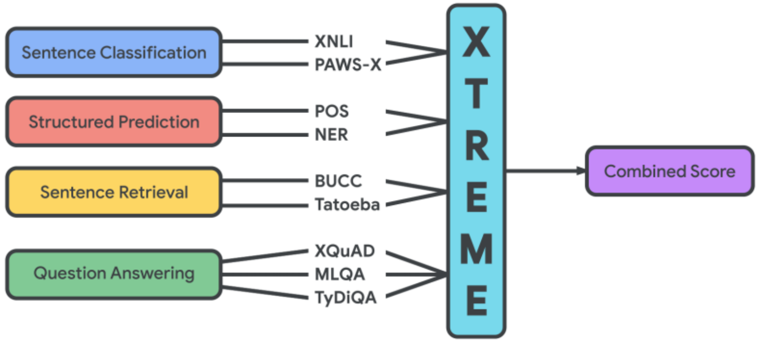
ID#	Feature Name	Category	Feature Values
1	Consonant Inventories	Phonology (19)	{1:Large, 2:Small, 3:Moderately Small, 4:Moderately Large, 5:Average}
23	Locus of Marking in the Clause	Morphology (10)	{1:Head, 2:None, 3:Dependent, 4:Double, 5:Other}
30	Number of Genders	Nominal Categories (28)	{1:Three, 2:None, 3:Two, 4:Four, 5:Five or More}
58	Obligatory Possessive Inflection	Nominal Syntax (7)	{1:Absent, 2:Exists}
66	The Perfect	Verbal Categories (16)	{1:None, 2:Other, 3:From 'finish' or 'already', 4:From Possessive}
81	Order of Subject, Object and Verb	Word Order (17)	{1:SVO, 2:SOV, 3:No Dominant Order, 4:VSO, 5:VOS, 6:OVS, 7:OSV}
121	Comparative Constructions	Simple Clauses (24)	{1:Conjoined, 2:Locational, 3:Particle, 4:Exceed}
125	Purpose Clauses	Complex Sentences (7)	{1:Balanced/deranked, 2:Deranked, 3:Balanced}
138	Tea	Lexicon (10)	{1:Other, 2:Derived from Sinitic 'cha', 3:Derived from Chinese 'te'}
140	Question Particles in Sign Languages	Sign Languages (2)	{1:None, 2:One, 3:More than one}
142	Para-Linguistic Usages of Clicks	Other (2)	{1:Logical meanings, 2:Affective meanings, 3:Other or none}

Example from Georgi, Xia and Lewis (2010)

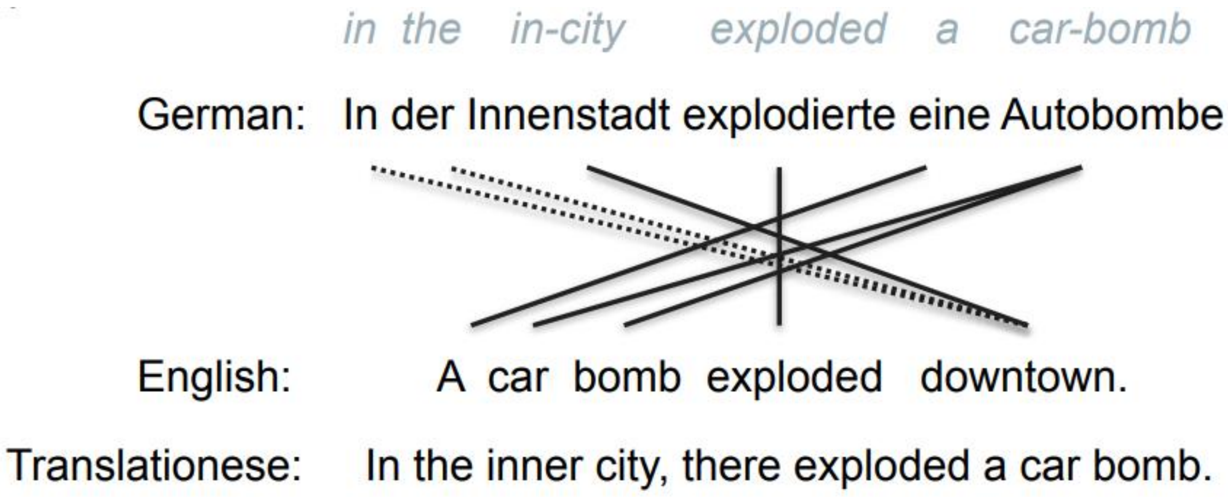
http://phontron.com/class/multiling2022/assets/slides/multiling-03-typology.pdf ; https://wals.info/

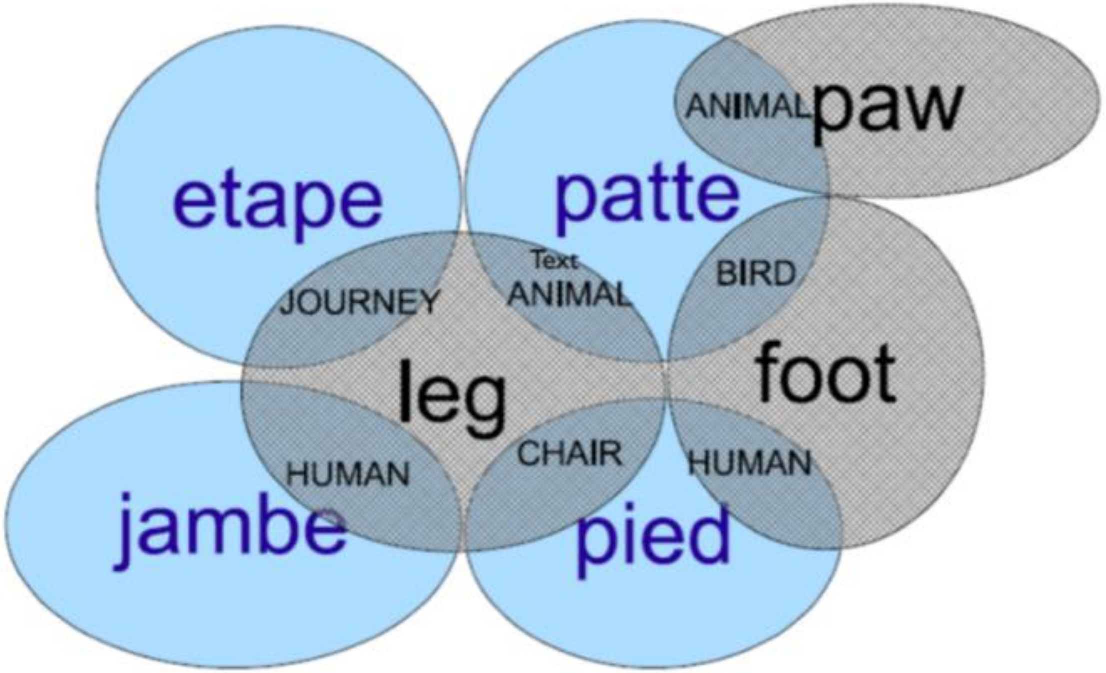
Examples of multilingual tasks

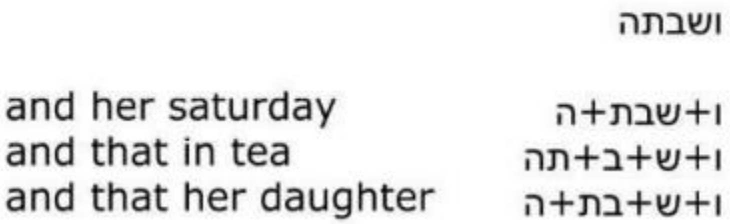
- Translation between multiple languages (e.g. FLORES)
- MASSIVE NLU benchmark in 51 language from Amazon Alexa
  - Recognize intents and slots in dialogues with assistant in any language
- NeuCLIR benchmark in cross-language information retrieval
  - Search among Zh, Fa and Ru documents with En queries
- Multilingual News Article Similarity
- Multilingual Complex Named Entity Recognition
- Composite benchmarks: XTREME, XGLUE



Why is it difficult to translate?







Эти типы стали есть на складе

- Материал находится на складе
- Люди едят на складе
- Сталь нужно есть на складе

Images source: https://web.stanford.edu/~jurafsky/slp3/

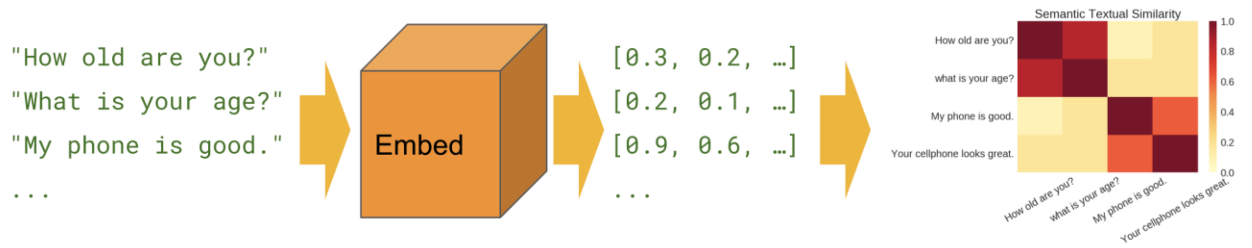
Examples of parallel corpora

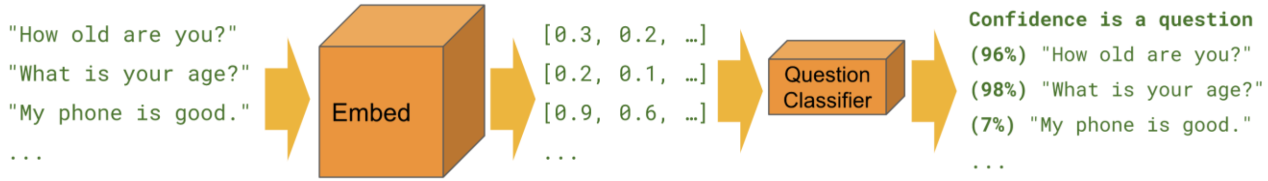
- Important books
  - Bible, Tanzil (Quran)
- Governmental texts
  - Europarl, UN corpus, etc.
- Subtitles
  - OpenSubtitles, TED, etc.
- Computer manuals
  - PHP, Ubuntu, etc.
- Aligned web data
  - ParaCrawl, WikiMatrix, CCMatrix, etc.
- A major repository: OPUS

Sentence encoders

Sentence encoders

- A sentence encoder is a model that converts a sentence (or another short text) to one fixed-size vector
  - BERT instead turns a text into a sequence of vectors (one per token)
    - But if we pool them (average or the embedding of [CLS]), it is a sentence encoder
- Applications:
  - Semantic similarity of sentences
    - Dot product or cosine similarity
    - Used e.g. for fast semantic search
  - Features for text classification
    - Works well in few-shot setting, especially with KNN
  - Cross-lingual transfer





Sources: https://tfhub.dev/google/universal-sentence-encoder/4; Cer et al, 2018, [Universal Sentence Encoder](#).

Training sentence encoders

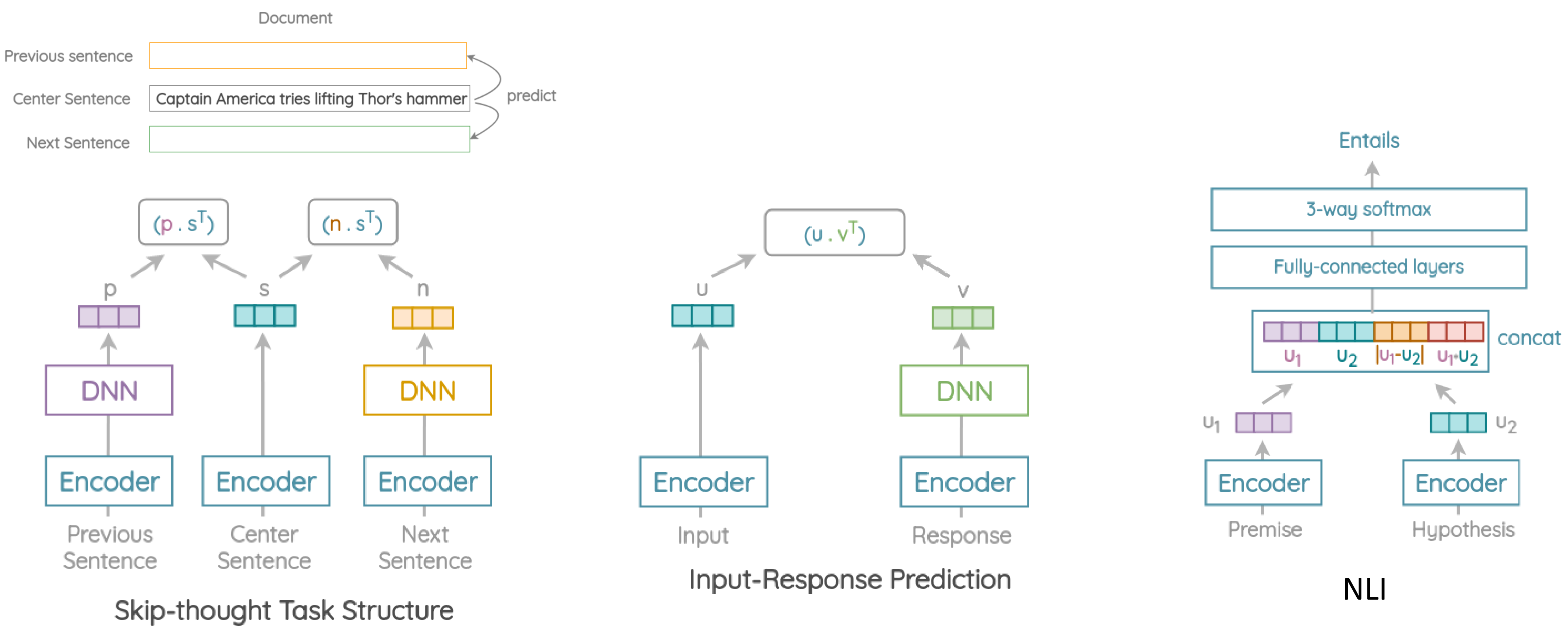
- Backbone model: transformer encoder[1,2,3,4,5], CNN[1,3]
- Pretraining: MLM[2,4,5], translation MLM[5]
- Training objectives:
  - Unsupervised contrastive learning[4]
    - Positive examples are created by augmentation (by simply applying dropout)
    - All other sentences in the batch are negative examples
  - Translation ranking[3,5]
    - Positive examples are sentences with the same meaning in another language
  - NLI[1,2,3,4]
    - entailment for positive examples, contradiction for negative examples
  - Skip-thought[1], predicting dialogue response[1,3], various classification tasks[1]

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \frac{e^{sim(x_i,y_i)/\tau}}{\sum_{j=1}^{N} e^{sim(x_i,y_j)/\tau}}$$

A typical loss for contrastive learning: softmax with (x_i, y_i) pairs as positive examples, (x_i, y_j) as negatives

1. Cer et al, 2018, Universal Sentence Encoder
2. Reimers et al, 2019, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks
3. Yang et al, 2019, Multilingual Universal Sentence Encoder for Semantic Retrieval
4. Gao et al, 2021, SimCSE: Simple Contrastive Learning of Sentence Embeddings
5. Feng et al, 2022, Language-agnostic BERT Sentence Embedding

Some pretraining tasks for USE



Source: https://amitness.com/2020/06/universal-sentence-encoder/

Evaluating sentence encoders

Sentence encoders can be evaluated on multiple tasks:

- Probing tasks: evaluate linguistic capacities
  - Predicting sentence length, dependency tree depth, word order, verb tense, etc.
- Downstream tasks: evaluate applications
  - Sentiment analysis (MR, CR, SST)
  - Subjectivity classification (SUBJ)
  - Opinion polarity (MPQA)
  - Question type detection (TREC)
  - NLI (SICK, SNLI)
  - Semantic similarity (STS, SICK)
  - Image-caption retrieval (COCO)

Model	MR	CR	SUBJ	MPQA	TREC	SST	MRPC
English Models							
InferSent	81.1	86.3	92.4	**90.2**	88.2	84.6	76.2
Skip-Thought LN	79.4	83.1	93.7	89.3	–	–	–
Quick-Thought	**82.4**	86.0	94.8	**90.2**	92.4	**87.6**	**76.9**
$\text{USE}_{\text{Trans}}$	82.2	84.2	**95.5**	88.1	93.2	83.7	–
Multilingual Models							
$m\text{-USE}_{\text{Trans}}$	78.1	**87.0**	92.1	89.9	**96.6**	80.9	–
LaBSE	79.1	86.7	93.6	89.6	92.6	83.8	74.4

A table from Feng et al, 2022, Language-agnostic BERT Sentence Embedding

- Sentence encoders are often used as fixed feature extractors
  - In contrast with other benchmarks such as SuperGLUE, where models are typically fine-tuned on each evaluation problem

Conneu et al, 2018, SentEval: An Evaluation Toolkit for Universal Sentence Representations

Some applications of sentence encoders

- Classifiers over fixed embeddings for few-shot classification or cross-lingual transfer
  - Efficient Intent Detection with Dual Sentence Encoders
  - Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic
  - EMET: Embeddings from Multilingual-Encoder Transformer for Fake News Detection
  - Deep Learning Models for Multilingual Hate Speech Detection
- Retrieval of relevant sentences and paragraphs
  - ReQA: An Evaluation for End-to-End Answer Retrieval Models
  - Intelligent Translation Memory Matching and Retrieval with Sentence Encoders
  - WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia

Text retrieval

Information retrieval

= Finding material that fulfills an information need
from within a large collection of unstructured documents.

Expression of Information Need	Potential Query	Potential Collection
Find related literature	The full text of the BERT paper	ACL anthology; arXiv CL
Recommend me a TV show to watch	[no explicit query!]	Netflix shows
Find every relevant patent	Boolean query with technical terms	U.S. Patents
Buy a new laptop	Short conversation: system asks questions to ascertain your criteria	E-commerce platforms

Source: http://web.stanford.edu/class/cs224u/

Non-neural search

- The problem:
  - Find (and rank) the most relevant documents related to the query
  - Do it fast!
- The basic solution: inverted index
  - Split the documents into words
  - Create the mapping from words to document ids
  - Consider only documents that contain the words from the query
  - Rank documents by TF-IDF, BM25, etc.
  - By remembering word positions in the document, we can look up by word n-grams

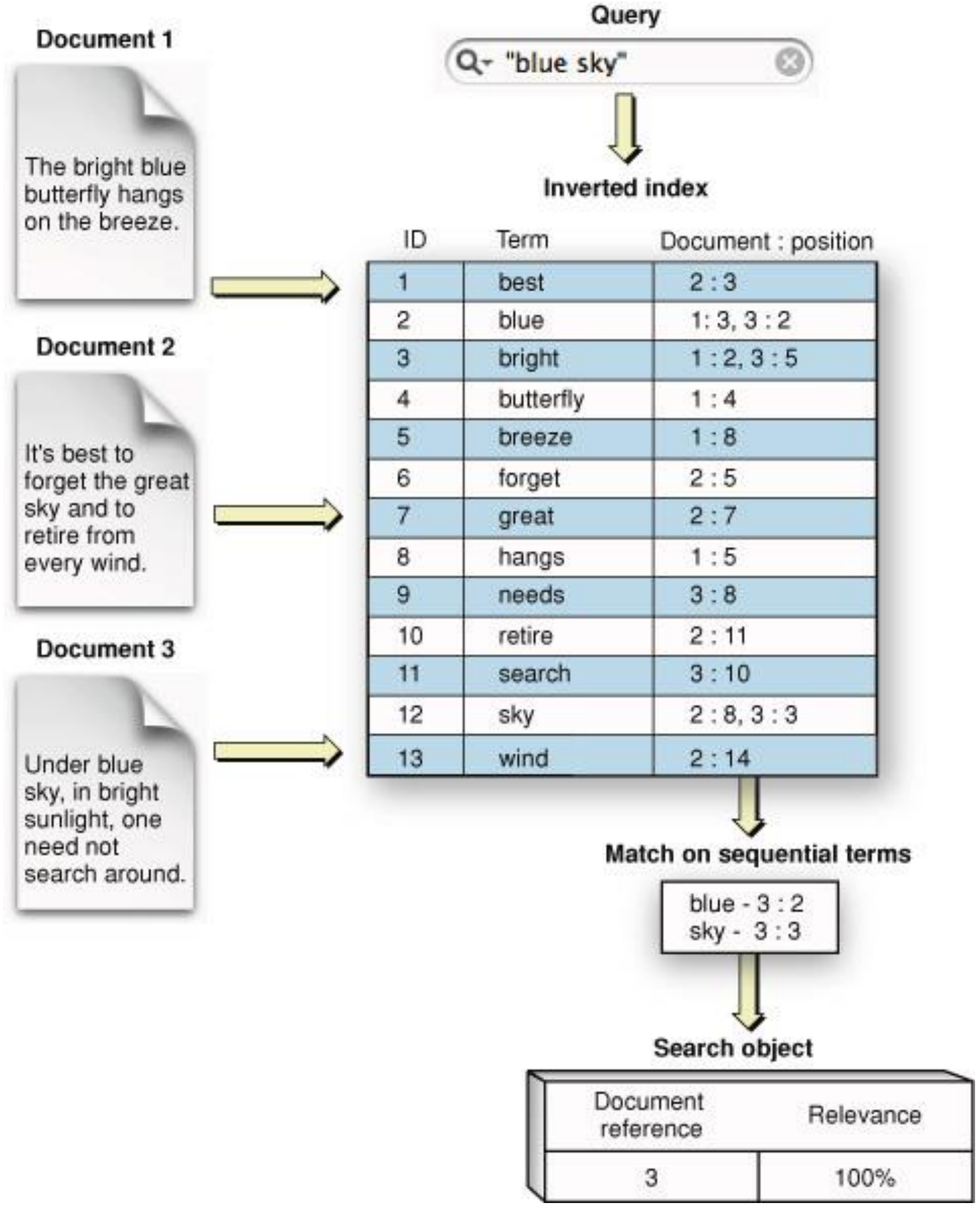


Image source: https://spec-zone.ru/RU/OSX/documentation/UserExperience/Conceptual/SearchKitConcepts/searchKit_basics/searchKit_basics.html

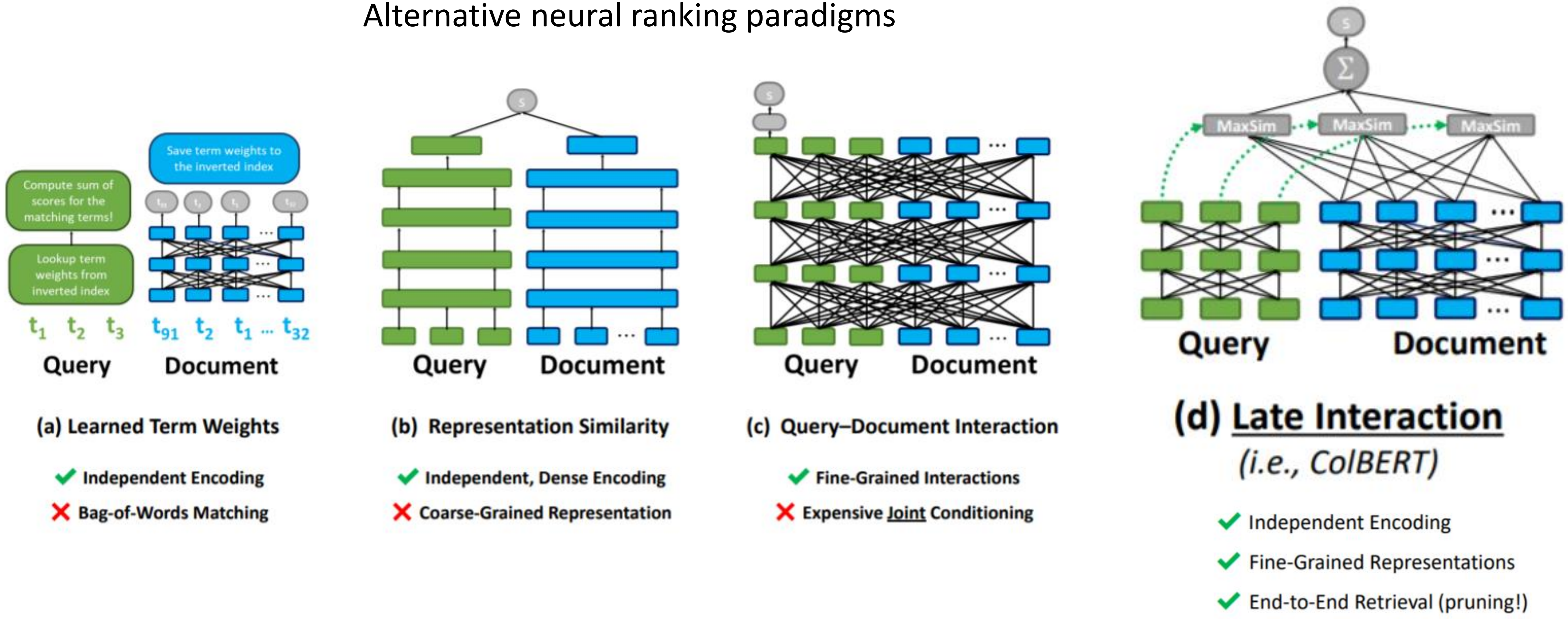
Transformers for text retrieval

- Retrievers:
  - Relevance = cosine_similarity(encoder(query), encoder(passage))
  - Passage embeddings can be precomputed once
  - Query and passage can be encoded with the same or different models (“Siamese network”)
  - Cosine similarity is equivalent to Euclidean distance → fast (approximate) nearest neighbor lookup is possible (e.g. FAISS)
- Rerankers:
  - Relevance = classifier(encoder(query + passage))
  - Works slower but better than retrievers due to cross-attention between query and passage
- The reranker can be combined with the reader
- The retriever or reranker can be trained jointly with the reader

Source: https://github.com/danqi/acl2020-openqa-tutorial/

Late document-query interaction

Alternative neural ranking paradigms



Sources: http://web.stanford.edu/class/cs224u/slides/CS224U-IR-part-5.pdf
https://arxiv.org/abs/2004.12832 (ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT)

Fast neighbor search

- Find the top k nearest neighbors among N d-dimensional vectors
- Linear search: O(N) (omitting k and d)
- Faster approaches:
  - KD tree, ball tree: use branching to eliminate the points far away, O(log(N))
- Approximations:
  - Locality-sensitive hashing: split space into buckets
  - HNSW: split space into hierarchy of buckets with increasing detail
- A popular implementation: FAISS
  - Example: match En and De Wikipedias in 3 hours (on 8 GPUs)

Multilingual models

How multilingual is multilingual BERT?

- The most typical multilingual pretrained models are BERT-like
  - E.g. multilingual BERT (2018), XLM(2019), XLM-R (2020), mDeBERTaV3 (2021)
- Most of them (except XLM) are fully unsupervised
- Still, they can perform cross-language transfer
- How does it even work???
  - Common vocabulary
  - Some mapping between vocabularies of similar languages
    - E.g. Hindi (Devanagari script) vs Urdu (Arabic script)
  - Generalization depends on the number of shared WALS features
- Perhaps, alignment occurs via shared words (e.g. URLs)

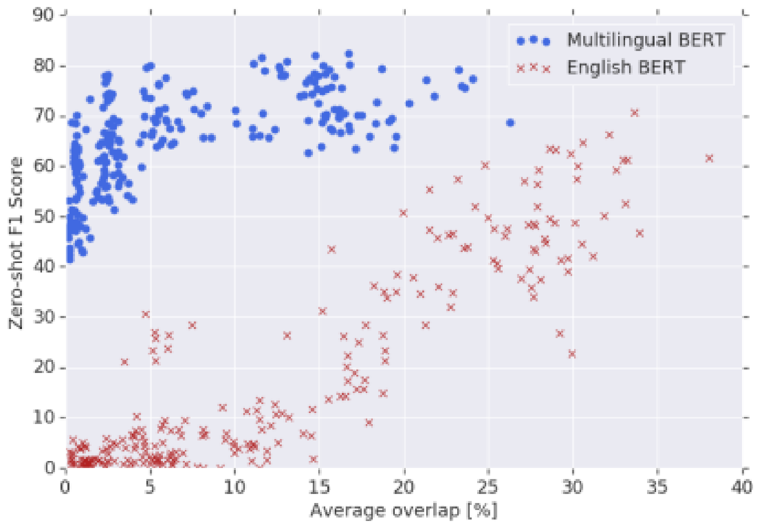


Figure 1: Zero-shot NER F1 score versus entity word piece overlap among 16 languages. While performance using EN-BERT depends directly on word piece overlap, M-BERT's performance is largely independent of overlap, indicating that it learns multilingual representations deeper than simple vocabulary memorization.

How multilingual is multilingual BERT? Pires et al, 2019, https://arxiv.org/abs/1906.01502

Multilingual generation

- XGLM by Meta
  - Pretrain a GPT-like model on a balanced corpus of 30 languages
  - Probe with few-shot in-context learning
    - SOTA in some generation tasks for lower-resourced languages
    - Capable of few-shot translation
- mGPT by Sber
  - Pretrained on 60 languages from Wikipedia and MC4
  - High scores in many zero-shot and few-shot tasks
- Both models can be fine-tuned for specific tasks

https://arxiv.org/abs/2112.10668 ; https://arxiv.org/abs/2204.07580

Multilingual seq2seq models

- [mT5](https://arxiv.org/abs/2010.11934)
  - Pretrained with a standard monolingual T5 denoising objective on mC4 (100 langs)
  - Fine-tuned for each task separately
  - SOTA on some multilingual NLI, NER and QA benchmarks
- [mBART](https://arxiv.org/abs/2001.08210)
  - Pretrained with a standard BART denoising objective on 25 languages; later extended to 50 languages
  - Language id is specified by the BOS token
  - Fine-tuned on translation pairs, achieved SOTA on low- and mid-resource languages
- [M2M100](https://arxiv.org/abs/2010.11125) and related models
  - An mBART-like transformer trained to translate between 2200 language pairs and 100 languages
  - The encoder produces nearly language-agnostic embeddings

https://arxiv.org/abs/2010.11934 ; https://arxiv.org/abs/2001.08210; https://arxiv.org/abs/2010.11125

Adapting BERTs to new languages

- Simplest: fine-tune the model on the target language
- Vocabulary adaptation: more efficient, but more complex
  - Remove unused tokens from the vocabulary (based on a target-lang corpus)
  - Add new tokens (e.g. by adding producing some BPE merges)
  - Initialize new embeddings using average embeddings of their constituents or source-language tokens aligned with them
  - Fine-tune the model on the target-lang (with e.g. MLM loss)
    - To speed it up, only embeddings can be fine-tuned (at least, for the 1st epoch)
- Training from scratch (which is more expensive)

https://aclanthology.org/W19-3712; https://arxiv.org/abs/1905.07213; https://aclanthology.org/2022.acl-short.68/

Tips for multilingual classification

- Augmentation with translated data helps
- Domain and task adaptation usually helps
- Multilanguage training usually helps
- Zero-shot transfer works OK, but worse than

Model	Data	DE	FR	JA	ES
multi-target	target	94.1	93.8	91.1	78.1
multi-all	all	93.8	94.3	91.4	77.7
zero-shot	EN	92.7	92.6	88.5	72.1

			CLS					HatEval				
Model	**Adapt.**	**Aug.**	**EN**	**DE**	**FR**	**JA**	**AVG**	**EN**	**EN†**	**ES**	**AVG**	**AVG†**
mono-target												
RoBERTa (EN) BERT (OTHERS)	×	×	$94.7_{0.4}$	$90.9_{0.6}$	$95.2_{0.0}$	$88.7_{0.3}$	92.4	$44.4_{5.3}$	$58.5_{6.2}$	$75.6_{0.6}$	60.0	67.1
		✓	**$95.3_{0.3}$**	$92.0_{0.2}$	$95.6_{0.3}$	$89.3_{0.02}$	93.0	$46.1_{2.6}$	$60.6_{3.2}$	$76.0_{1.7}$	61.0	68.3
	TAPT	×	$94.9_{0.1}$	$91.6_{0.1}$	$95.4_{0.1}$	$89.3_{0.3}$	92.8	$45.4_{1.9}$	$59.9_{2.7}$	$76.1_{1.1}$	60.8	68.0
		✓	$95.0_{0.4}$	$92.3_{0.4}$	$95.8_{0.2}$	$89.7_{0.4}$	93.2	$44.7_{1.5}$	$59.2_{1.7}$	$76.9_{1.4}$	60.8	68.0
	TAPT+ DAPT	×	$94.9_{0.4}$	$91.8_{0.2}$	$95.5_{0.3}$	$89.5_{0.2}$	92.9	$48.0_{1.5}$	$63.1_{2.6}$	$76.3_{1.1}$	62.2	69.7
		✓	**$95.3_{0.1}$**	$93.0_{0.8}$	**$95.9_{0.1}$**	$89.9_{0.4}$	**93.5**	$46.0_{4.3}$	$60.2_{4.4}$	$76.9_{0.6}$	61.4	68.5
multi-target												
XLM-RoBERTa	×	×	$92.5_{0.4}$	$93.0_{0.2}$	$92.5_{0.3}$	$90.4_{0.5}$	92.1	$47.2_{2.0}$	$61.4_{1.9}$	$74.8_{0.5}$	61.0	68.1
		✓	$93.3_{0.1}$	$94.0_{0.2}$	$93.8_{0.2}$	$90.3_{0.3}$	92.8	$45.6_{1.6}$	$59.3_{2.5}$	$77.0_{1.1}$	61.3	68.1
	TAPT	×	$92.7_{0.5}$	$93.5_{0.5}$	$93.9_{0.3}$	$90.3_{0.1}$	92.6	$47.0_{2.7}$	$62.4_{3.3}$	$76.1_{1.4}$	61.6	69.2
		✓	$93.4_{0.6}$	$94.0_{0.3}$	$93.8_{0.5}$	$90.5_{0.4}$	92.9	$47.9_{1.3}$	$63.5_{1.5}$	$77.9_{0.9}$	62.9	70.7
	TAPT+ DAPT	×	$93.1_{0.6}$	$93.0_{0.5}$	$93.6_{0.1}$	$90.8_{0.3}$	92.6	$49.9_{2.5}$	$65.6_{2.4}$	$76.5_{1.0}$	63.2	71.0
		✓	$94.0_{0.3}$	**$94.1_{0.4}$**	$93.8_{0.3}$	$91.1_{0.4}$	93.2	$46.6_{2.1}$	$61.7_{2.5}$	**$78.1_{0.8}$**	62.3	69.9
multi-all												
XLM-RoBERTa	×	×	$92.4_{0.3}$	$92.6_{0.4}$	$93.3_{0.4}$	$90.4_{0.4}$	92.2	$48.4_{3.5}$	$63.1_{4.5}$	$77.5_{0.4}$	62.9	70.3
		✓	$93.4_{0.3}$	$93.3_{0.2}$	$94.0_{0.2}$	$90.4_{0.5}$	92.8	$49.8_{3.5}$	$66.0_{4.6}$	$77.8_{0.9}$	63.8	71.9
	TAPT	×	$92.5_{0.4}$	$93.0_{0.3}$	$93.9_{0.3}$	$90.9_{0.3}$	92.6	$48.4_{2.7}$	$64.2_{3.5}$	$77.4_{0.9}$	62.9	70.8
		✓	$93.5_{0.4}$	$93.4_{0.5}$	$94.1_{0.2}$	$91.1_{0.2}$	93.0	$50.0_{2.2}$	$66.5_{2.6}$	$77.8_{0.6}$	63.9	72.2
	TAPT+ DAPT	×	$92.7_{0.3}$	$93.3_{0.2}$	$94.0_{0.3}$	$91.2_{0.3}$	92.8	$47.1_{3.9}$	$62.7_{5.3}$	$77.4_{1.0}$	62.3	70.1
		✓	$93.5_{0.3}$	$93.8_{0.2}$	$94.3_{0.3}$	**$91.4_{0.2}$**	93.3	**$50.7_{1.1}$**	**$67.4_{1.4}$**	$77.7_{0.7}$	**64.2**	**72.6**

Practical Transformer-based Multilingual Text Classification, Wang and Banco, 2021: https://aclanthology.org/2021.naacl-industry.16

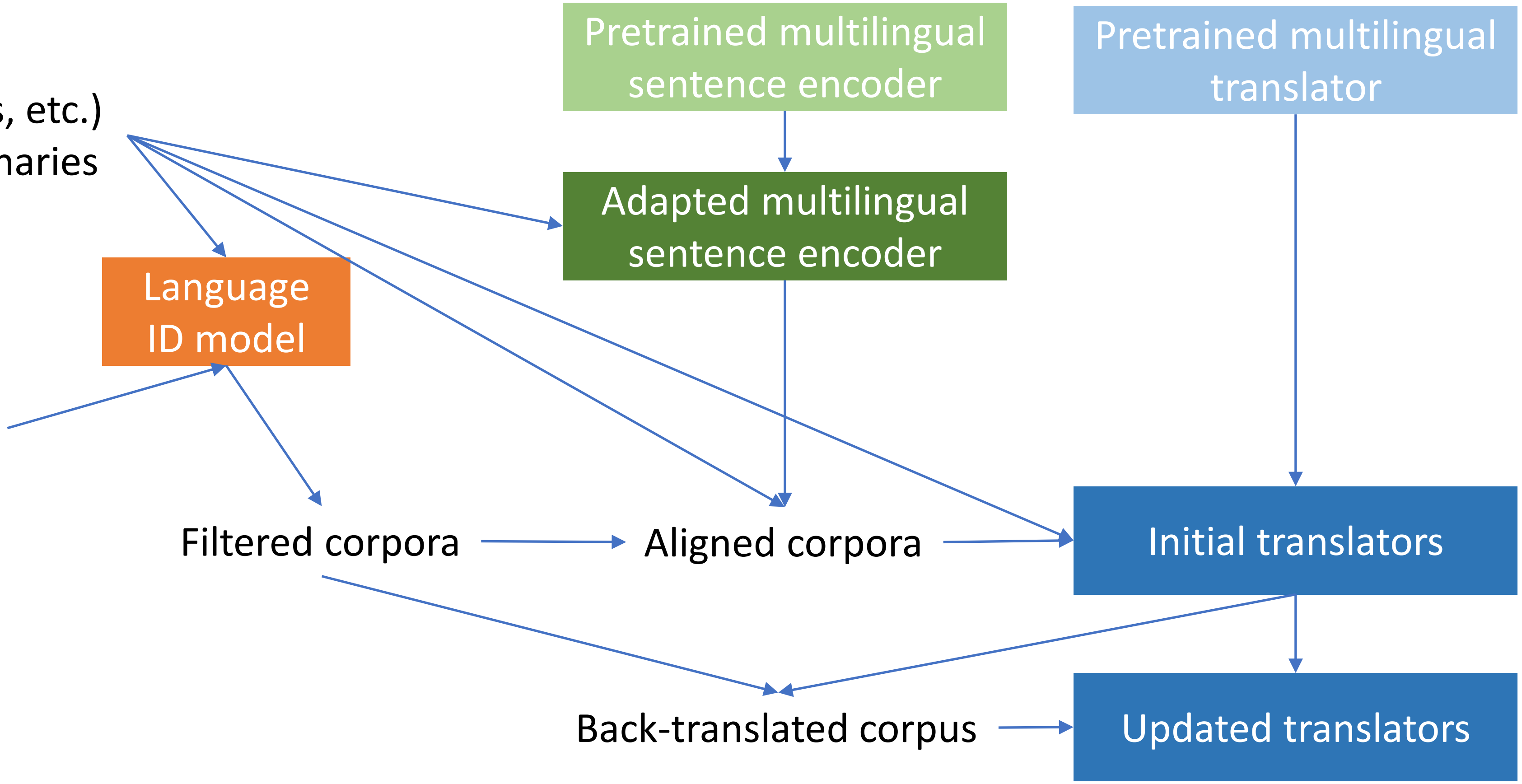
How to bootstrap NLP for a new language?

Typical initial resources:

- Small parallel data (bible, laws, etc.)
- Word- and phrase-level dictionaries
- Wikipedia

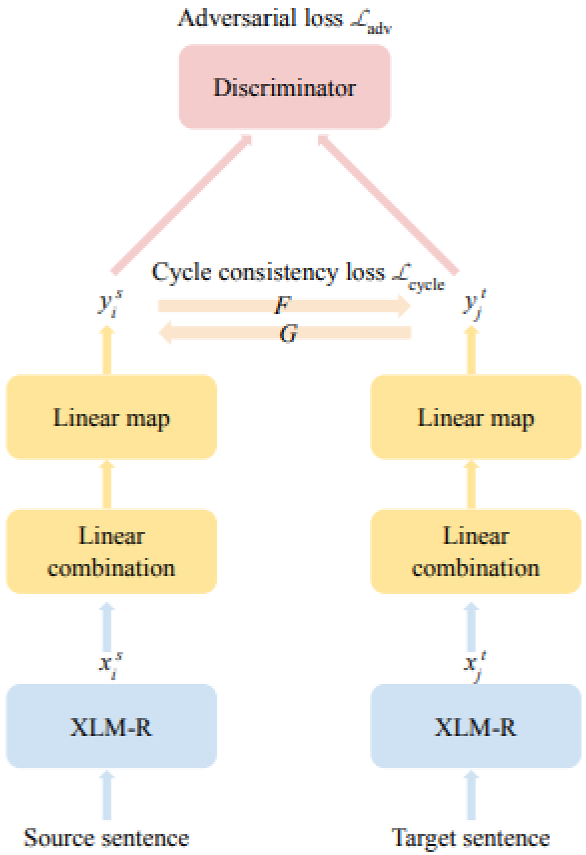
Dirty corpora

- Wikipedia in other languages
- Mixed-language web crawl
- Unaligned parallel literature



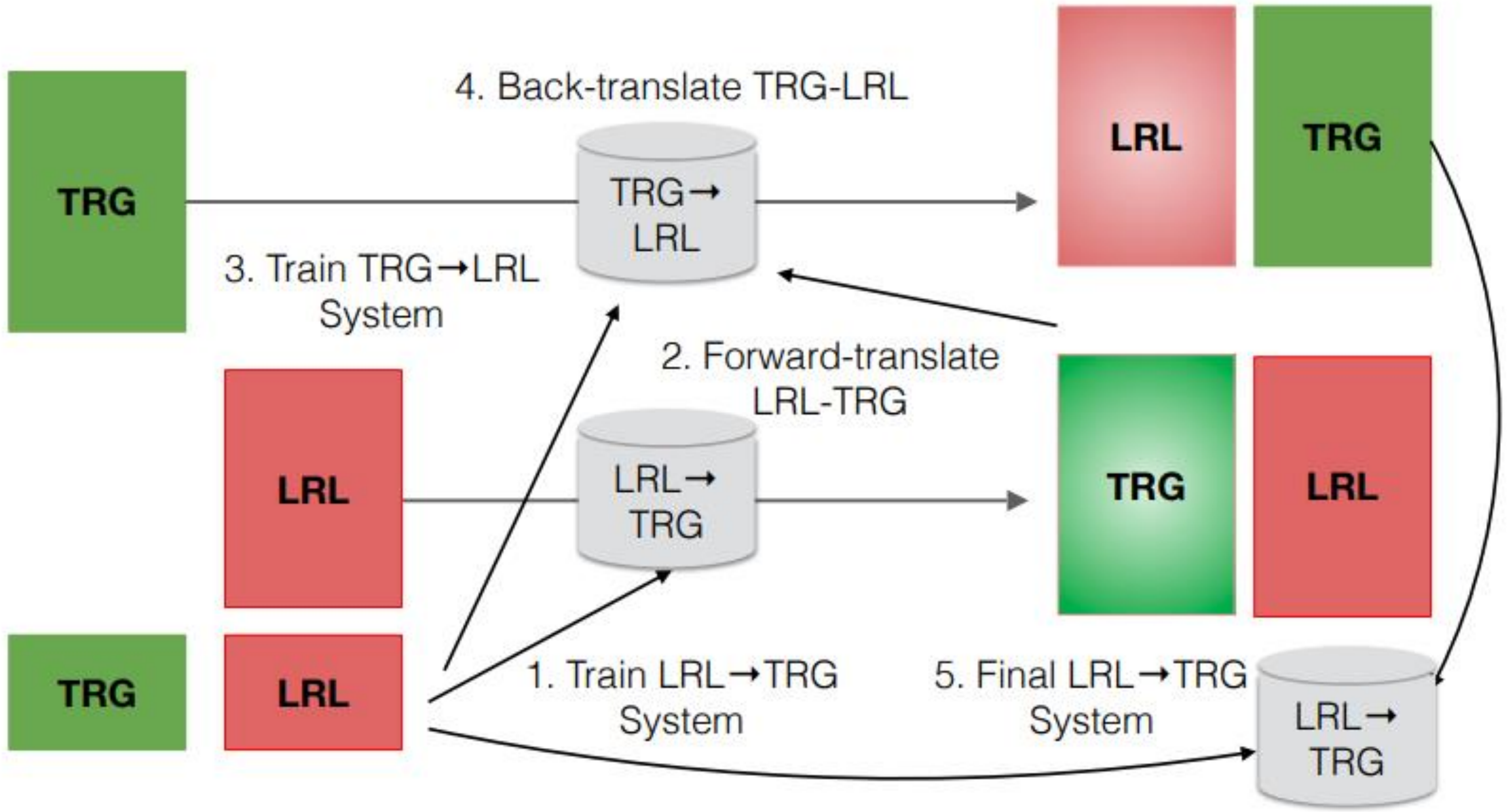
XLM-R –> universal sentence encoder

- Use XLM-R as a fixed feature extractor
- Train a weighed average pooler and FFN head to extract cross-lingual embeddings
  - It is possible to train it even an unsupervised way: with cycle consistency and adversarial loss
- Such a model can be used for matching sentences in unseen languages
  - (Because XLM-R has already seen them)



https://arxiv.org/abs/2104.07642

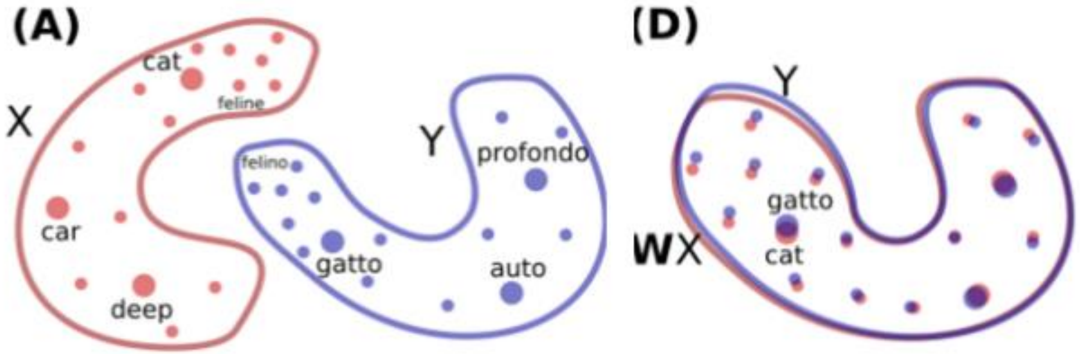
Iterative back-translation



Source: http://phontron.com/class/multiling2022/assets/slides/multiling-07-mtaugmentation.pdf

Unsupervised word translation

- Hypothesis: Word embedding spaces in two languages are isomorphic
  - One embedding space can be linearly transformed into another
  - Give monolingual embeddings X and Y, learn a (orthogonal) matrix, such that, WX = Y
- Use adversarial learning to learn W:
  - If WX and Y are perfectly aligned, a discriminator shouldn't be able to tell
  - Discriminator: Predict whether an embedding is from Y or the transformed space WX.
  - Train W to confuse the discriminator



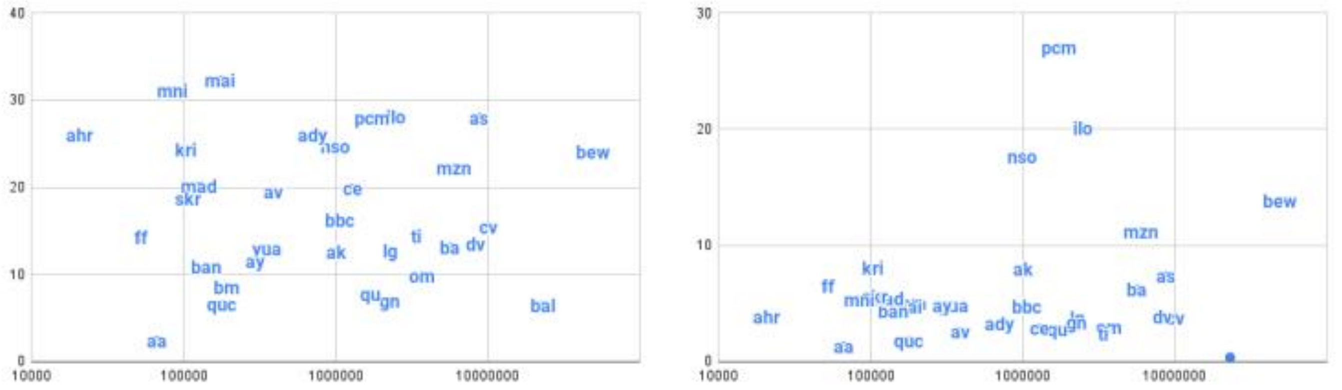
After aligning words, a sentence translation model can be trained:

- Pretrain with monolingual denoising
- Finetune with back-translation

http://phontron.com/class/multiling2022/assets/slides/multiling-10-unsupervisedmt.pdf

What is next?

- *Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning*, a recent paper by Google
- The bootstrapping pipeline for 1000 languages
  - Language identification
  - Monolingual denoising pretraining in all languages
  - Fine-tuning on en->xx and xx->en pairs
  - Good translation for some zero-resource languages



(a) Any-to-English (xx → en)

(b) English-to-Any (en → xx)

Figure 2: Unsupervised/zero-resource BLEU on 30 new languages. The x-axis depicts the amount of monolingual data available for the language, while the y-axis depicts the BLEU score of the 1.6B parameter Transformer model after fine-tuning with online back-translation. The data point corresponding to each language is represented by its BCP-47 language code.

https://arxiv.org/abs/2201.03110

Conclusions

- Multilanguage NLP is difficult and important
- Multilingual sentence encoders are an important resource
- There are multilingual encoder, decoder, and enc+dec transformers
- NLP resources for new languages can be bootstrapped