Advanced Methods in Text Analytics

Multilingual NLP

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Most Spoken Languages (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- What are the most spoken languages in the world?
  - According to Ethnologue, this is the distribution for *native* speakers

Languages with at least 50 million first-language speakers[7]

	Language	Native speakers (in millions)	Language family	Branch
1	Mandarin Chinese	941	Sino-Tibetan	Sinitic
2	Spanish	486	Indo-European	Romance
3	English	380	Indo-European	Germanic
4	Hindi	345	Indo-European	Indo-Aryan
5	Bengali	237	Indo-European	Indo-Aryan
6	Portuguese	236	Indo-European	Romance
7	Russian	148	Indo-European	Balto-Slavic
8	Japanese	123	Japonic	Japanese
9	Yue Chinese	86	Sino-Tibetan	Sinitic
10	Vietnamese	85	Austroasiatic	Vietic

Image source

- The top 10 languages cover 3 billion people (37.5% of the world)

Most Spoken Languages (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- What if we break this into *native vs non-native* speakers?
  - Again, according to Ethnologue...

Most spoken languages, *Ethnologue*, 2023[4]

	Language	Family	Branch	First-language (L1) speakers	Second-language (L2) speakers	Total speakers (L1+L2)
1	English (excl. creole languages)	Indo-European	Germanic	380 million	1.077 billion[5]	1.456 billion
2	Mandarin Chinese (incl. Standard Chinese, but excl. other varieties)	Sino-Tibetan	Sinitic	939 million	199 million[6]	1.138 billion
3	Hindi (excl. Urdu)	Indo-European	Indo-Aryan	345 million	266 million[7]	610 million
4	Spanish (excl. creole languages)	Indo-European	Romance	485 million	74 million[8]	559 million
5	French (excl. creole languages)	Indo-European	Romance	81 million	229 million[9]	310 million

Image source

- The top 5 languages cover 4 billion people (50% of the world)

Most Spoken Languages (3)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- What about language distribution as used on the *world wide web*?
  - According to W3 Technology Surveys...

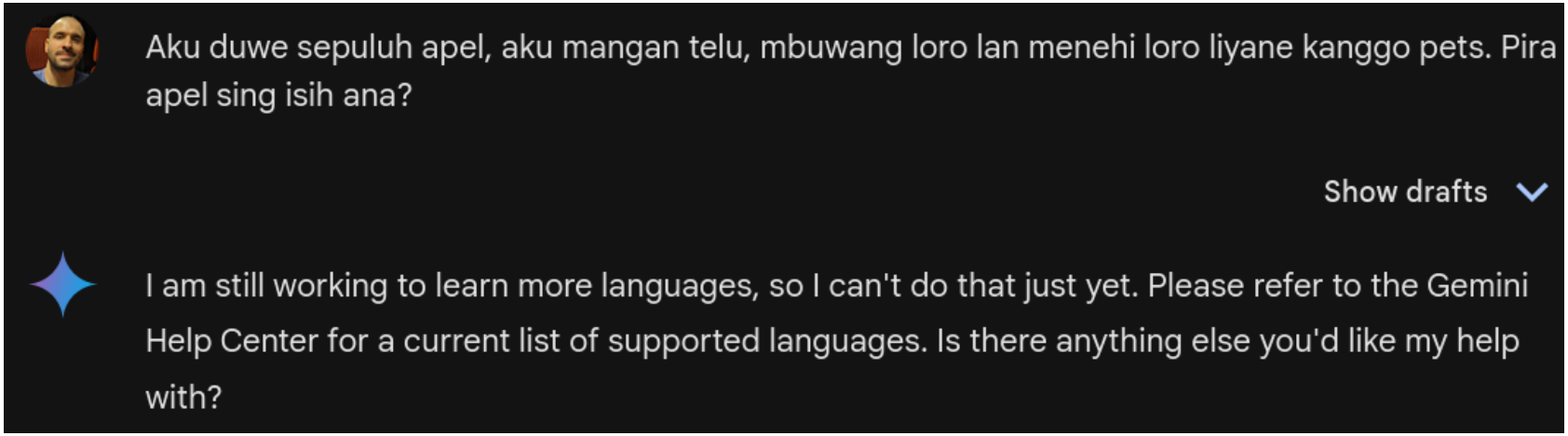
Rank	Language	16 May 2023	07 May 2024
1	English	55.5%	50.5%
2	Spanish	5.0%	5.7%
3	Russian	4.9%	4.2%
4	German	4.3%	5.2%
5	French	4.4%	4.3%
6	Japanese	3.7%	4.7%
7	Portuguese	2.4%	3.5%
8	Turkish	2.3%	1.9%
9	Italian	1.9%	2.5%
10	Persian	1.8%	1.4%

- English alone covers 50% of the content online (top 10 languages 84%)
  - **Massive difference in language distribution in the world vs online**

Why Multilingual NLP? (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **NLP:** roughly, automatically processing *human language* with computers
  - There are over *7100 human languages* in the world (source)
  - **Ideally**, all of these languages have **equal access to information**

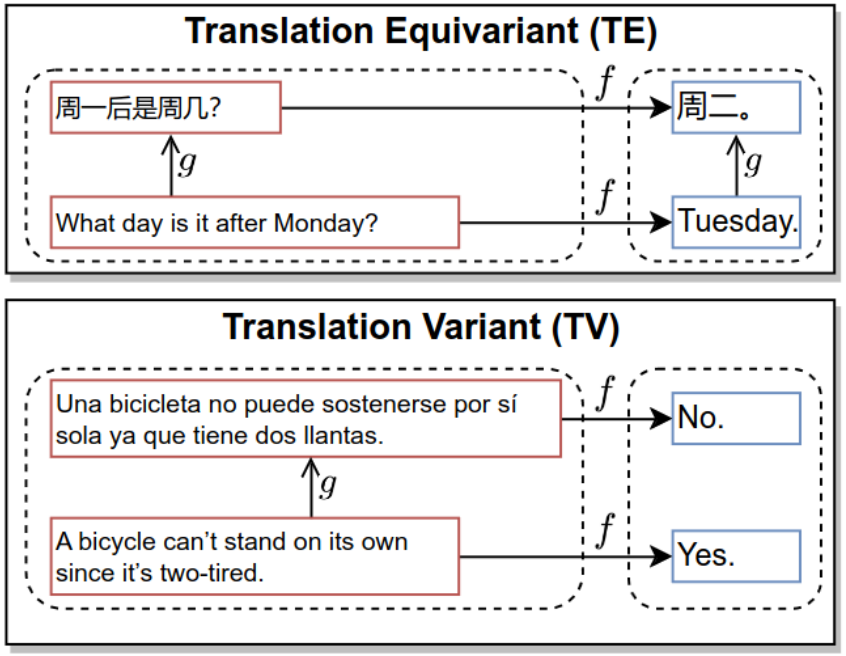


- Google Gemini (released Dec. 2023) does not speak *Javanese*
  - But 68 million people speak this language in Indonesia
  - That's the same number of people that speak Italian
- Why this limitation then?

Why Multilingual NLP? (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **State-of-the-art** (SOTA) **NLP** methods are **statistical methods**
  - In other words, lots of data required to achieve SOTA performance
  - E.g. large language models (LLMs) typically trained on a massive corpus crawled from the web
- **So, performance in English much better than in less available languages**
  - **Translation not always a solution**
  - E.g. if output is independent of input language, LLMs can solve problems via translation (Zhang et al. 2023)
  - But not if input language changes answer
  - Example in bottom: pun detection



Why Multilingual NLP? (3)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Multilingual natural language processing (mNLP):** NLP systems that can process *multiple* natural languages
  - But how many languages?
  - Ideally, the more the better!
  - But, the more languages, the more challenges
- **In this lecture:**
  - Cross-lingual transfer (fundamental goal of mNLP)
  - Multilingual transformers

Outline

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

1. Cross-Lingual Transfer

2. Multilingual Transformers

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Cross-Lingual Transfer

What is Cross-Lingual Transfer? (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Due to their statistical nature, SOTA NLP models perform much better in English than in languages less available online (Zhang et al. 2023)
  - E.g. GPT-3 (Brown et al. 2020) and LlaMA-2 (Touvron et al. 2023) trained on data that is heavily skewed toward English
  - Supervised datasets for downstream fine-tuning mostly available in English
- **Cross-lingual transfer:** use knowledge a model obtained in some *source language* in an application in a different *target language*
  - **Source language *s*** usually a high resource language, e.g. English
  - **Target language *t*** usually a low-resource language, e.g. Quechua
  - Essentially, **transfer learning across languages**
- General example:
  - Pre-train or fine-tune LM on data in English (*source* language), then fine-tune LM or do inference with it using data in Portuguese (*target language*)
- Why would you do that? More data -> stronger "source model"
  - **Is what the "source model" has learned transferable to other languages?**

What is Cross-Lingual Transfer? (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- More common setting:
  - **First pre-train** LM on English and Javanese, e.g. a CLM or MLM.
  - **Then fine-tune** on some specific task, e.g. sentiment analysis, using data in English (safe to assume lots of data for the task in English)
  - **Then do inference** on same task using data in Javanese (where there is little data in this language for fine-tuning your LM for the task)
- Easier said than done, more details to handle in most cases
- Success is cross-lingual transfer suggests:
  - There are **aspects of the task** that are **shared across languages**
  - Trained models learned these shared/transferable aspects of the task
- This success could enable **democratization of SOTA NLP methods!**
  - Thus, cross-lingual transfer is a **fundamental goal** of mNLP
- How to get there? Normally, by relying on **shared cross-lingual features**
  - I.e. features that are similar across more than one language
  - Easy way to visualize this: cross-lingual word embeddings (**CLWE**)

Embedding Space Alignment (1)

- Say we **train static word embeddings** with skip-gram (Mikolov et al. 2013) **in two different languages**: English and Italian
  - The embedding space may look as shown below (clustered by language)

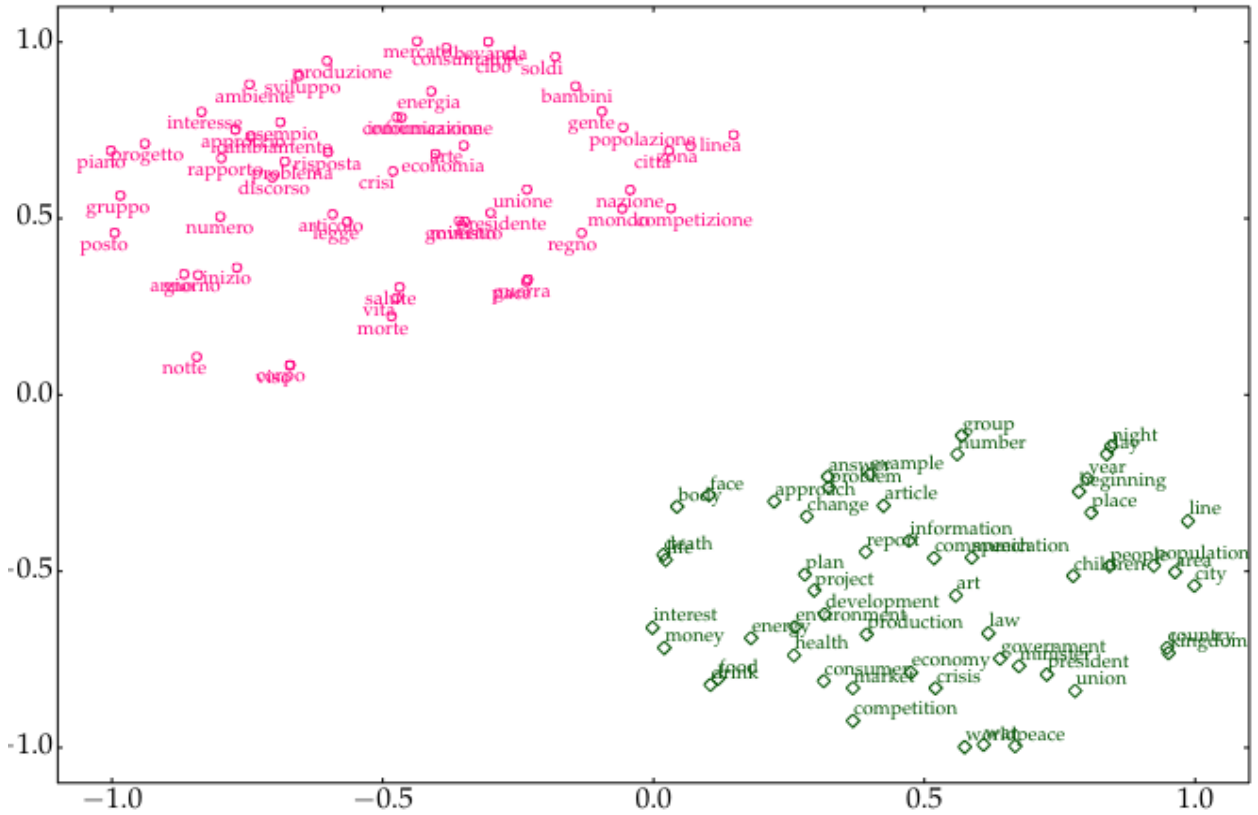


Image source

Embedding Space Alignment (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- We want **shared features across languages**, i.e. a bilingual embedding space clustered (often referred to as *aligned*) by meaning
  - This requires an **alignment signal**, e.g. word-to-word translations

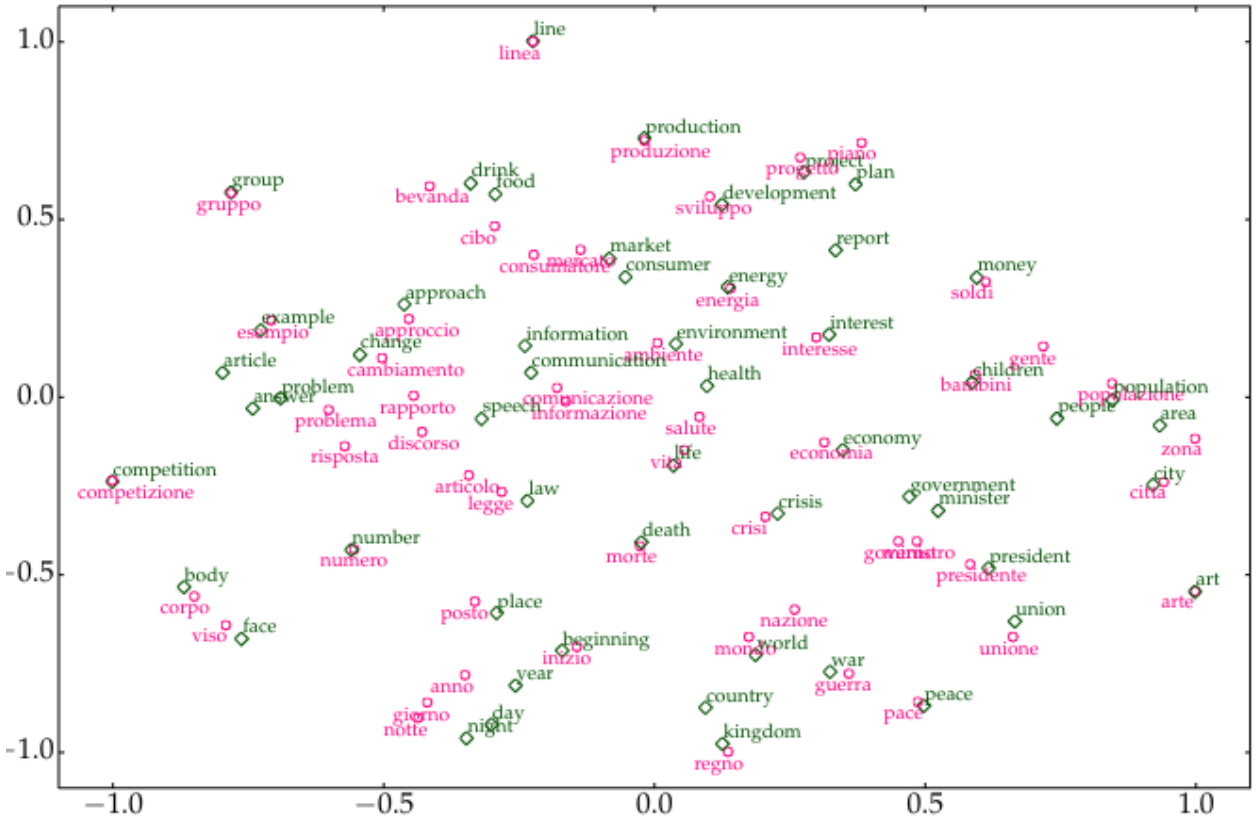
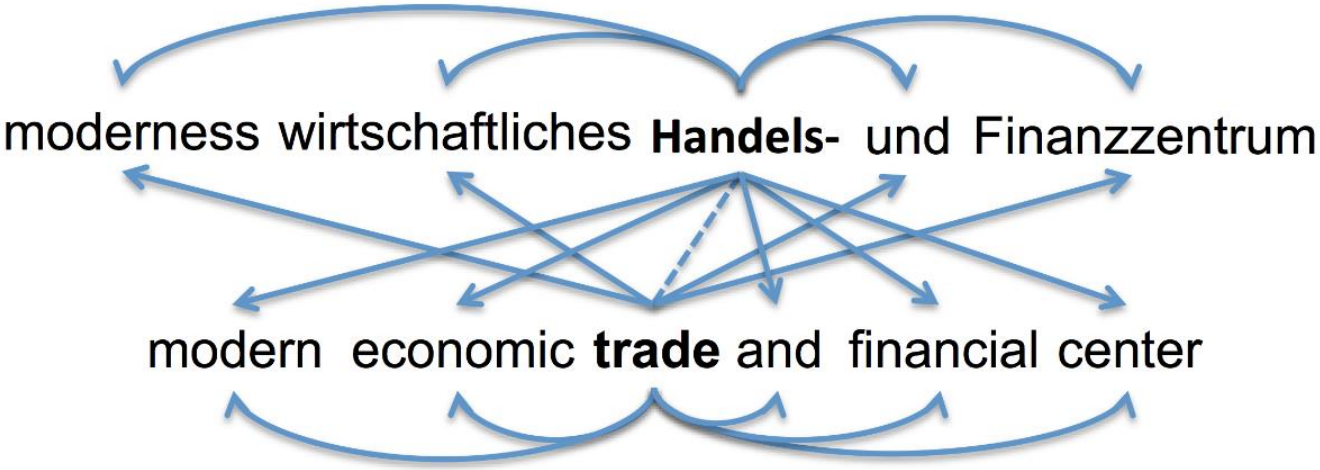


Image source

Cross-Lingual Word Embeddings (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Example:** the bilingual skip-gram model by Luong et al. 2015
  - Recall **skip-gram:** given center word, predict neighboring words
- **Bilingual skip-gram:** in addition to *monolingual* objective (curved arrows), predict surrounding words in target language (straight arrows)



- Given source language *s* and target language *t*, **such a method requires:**
  - Corpora C_s, C_t, vocabularies V_s, V_t and parameters $\boldsymbol{W}_1^s$, $\boldsymbol{W}_2^s$, $\boldsymbol{W}_1^t$, $\boldsymbol{W}_2^t$
  - Dictionary $D_s = \{(s_i^s, s_i^t)\}$ of *sentence translation pairs* where $s_i^s \in V_s$, $s_i^t \in V_t$

Cross-Lingual Word Embeddings (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Why does this method require *parallel sentences*?
  - I.e. why do we require translation?
  - And why translations of sentences and not just words?
- **Parallel sentences**, i.e. translated sentences, is **alignment signal** to enforce shared features across languages
  - **Task dependent:** for skip-gram objective, which is based on predicting context words given center word, we require **sentences** to **provide suitable context windows during training** (principle of **distributed semantics**)
  - Such parallel sentences commonly found in machine translation datasets
- But **how do we use *D* to align** corresponding words in $\boldsymbol{W}_i^s$ and $\boldsymbol{W}_i^t$?
  - We could add a term to the training objective to force model to minimize distance between corresponding vectors for *aligned words* in $(s_i^s, s_i^t) \in D_s$
- What did the authors do?
  - Set $\boldsymbol{w}_i^s = \boldsymbol{w}_j^t$ for *aligned words* in every example from D_s, where $\boldsymbol{w}_i^s$, $\boldsymbol{w}_j^t$ are i-th and j-th row of $\boldsymbol{W}_i^s$ and $\boldsymbol{W}_i^t$, resp. (additional word alignment needed)

Evaluating CLWE

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Intrinsic evaluation:**
  - Bilingual lexicon induction (BLI)
  - Cross-lingual word similarity (XL-SIM)
- **BLI:** given $(w_i^s, w_i^t) \in D$, take $\boldsymbol{w}_i^s$ as query and rank all rows from $\boldsymbol{W}^t$ based on similarity to $\boldsymbol{w}_i^s$, then check rank r_i of correct answer $\boldsymbol{w}_i^t$
  - We compute metrics based on r_i averaged over all w_i
  - E.g. precision@1 (P@1): percentage of k pairs where $r_i = 1$
  - Mean reciprocal rank (MRR): average of $1/r_i$ across k pairs
- **XL-SIM:** given $(w_i^s, w_i^t) \in D$, compare similarities of corresponding vectors with semantic similarity *scores* given by human annotators
  - Subjective task, requires averaged scores across multiple annotators
- **Extrinsic evaluation:** use CLWE on **downstream cross-lingual tasks**
  - Zou et al. 2013 used CLWE as features in phrase-based **machine translation**
  - **Entity linking:** match *mention* in one language to *entity* in another language

Beyond CLWE

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Static representations** of words seldom used today
  - Instead, NLP relies almost exclusively on **contextualized representations**
  - Typically, these representations are provided by transformer-based language models (LMs)
  - Even if multilingual, static representations are still limited
- Nevertheless, the **following concepts** are **still relevant in mNLP**:
  - **Cross-lingual transfer**
  - **Alignment of representation space**
- In the next section:
  - Cross-lingual transfer using transformer-based models

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Multilingual Transformers

Multilingual BERT

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- What could cross-lingual transfer look like with transformer-based LMs?
  - Any ideas?
- Pires et al. 2019 developed multilingual BERT to explore this question
  - They framed **cross-lingual transfer with BERT** as follows
    1. Pre-Train BERT on corpus of concatenation of text (Wikipedia) in multiple different languages (let's call it **mBERT**)
    2. Fine-tune mBERT on some task using language *s* (seen in pre-training)
    3. Evaluate mBERT on same task using different language *t* (seen in pre-training)
  - They called this ***zero-shot cross-lingual transfer***
  - Note: **no cross-lingual supervision**, e.g. parallel sentences
- **If successful**, it would mean the **model learns the task beyond its representation in some language**
  - "Learns the task", some aspects of it at least
  - **Example task:** using an ATM. Some aspects of it are beyond language, e.g. button placement, order of actions (first insert card, etc.)

Training Multilingual LMs

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Training data was highly unbalanced in mBERT
  - This has important implications on performance
- **Training corpus:** concatenation of Wikipedia in 104 different languages
  - But size of Wikipedia is different in different languages
  - Number of articles in English: about 6.8M articles
  - Number of articles in Vietnamese: about 1.3M articles
  - Number of articles in Javanese: about 73000 articles
- **Impact on performance** is predictable
  - Performance on more common languages is better. Why?
  - Transformer has more data to *learn* from
  - *Learn* = adjust its 100M parameters so it generalizes instead of overfitting
- In addition, **low-resource languages typically require more data given same number of parameters**
  - Why?
  - **Hint:** the model uses a **single shared subword vocabulary** for all languages

Tokenization in Multilingual LMs (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **What impact does unbalanced data across languages have on tokenization?**
  - mBERT used WordPiece (mentioned in tokenization lecture)
  - **Recall WordPiece:** similar to BPE (covered in detail in tokenization lecture)
    1. Start with vocabulary *V* made of characters
    2. Merge pairs of tokens that increase data likelihood of n-gram LM
    3. Keep merging until you reach desired vocabulary size
  - If we add frequent tokens to vocabulary, a count-based n-gram model can better estimate their probabilities.
  - Similar effects in other forms of tokenization, e.g. BPE, UnigramLM
- In other words, **multilingual vocabulary dominated by "larger" languages, as measured by training data size**
  - Recall example from tokenization lecture: *hello* in English -> 1 token
  - Corresponding word in Hindi: *namaste* -> 3 tokens
- Why does this matter?

Tokenization in Multilingual LMs (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Impacts of multilingual tokenization already discussed in tokenization lecture
  - **Same message requires more tokens** to be represented **in low-resource languages**
  - Given limited input length in modern LMs, **low-resource languages are limited to shorter input sequences**
- But **why does this mean low-resource languages require more data** given the same number of parameters?
  - Because transformer needs to **learn to contextualize subword tokens that belong to the same word**
  - E.g. learn that *na*, *ma* and *ste* should attend to one another
  - Can be challenging, because **subword tokens more likely to appear in different languages, but with different meanings** (e.g. article *a* in English)
- Note **vicious circle**: low-resource languages train with low amounts of data, which creates problems that can be solved with more data

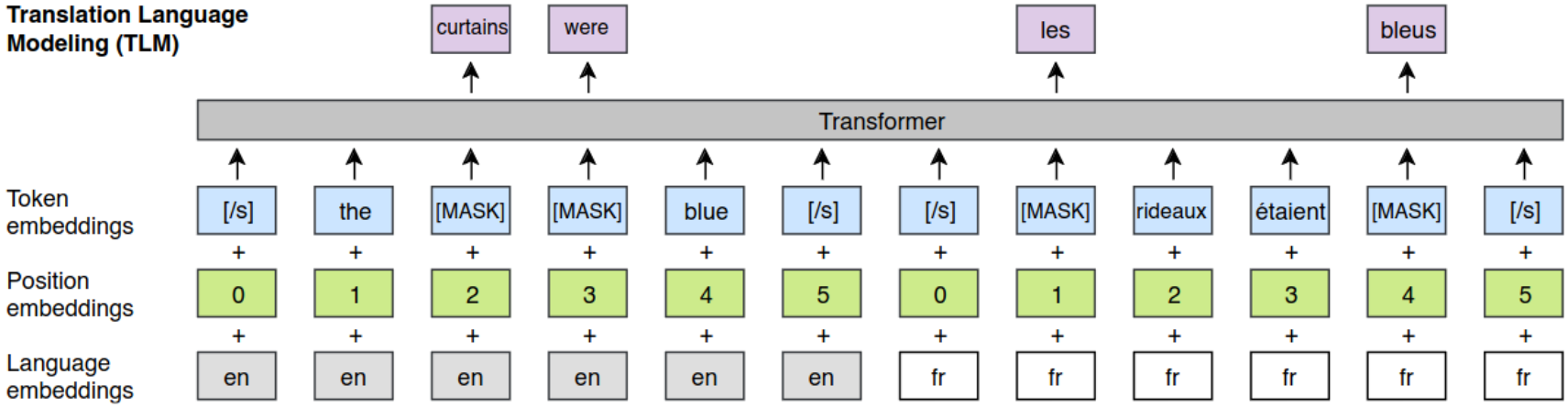
Generalizing Across Languages (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- mBERT was indeed "surprisingly" successful despite unbalanced data
  - I.e. the model was quite good at zero-shot cross-lingual transfer
  - This suggests model can do tasks beyond language
  - Why surprising?
- **How can we explain mBERTs success without alignment signal?**
  - Recall: **no alignment signal**, e.g. parallel sentence translations
- Several studies have focused on this question and found:
  - K et al. 2020: shared vocabulary not needed, model depth needed, languages should be of similar structure
  - Artetxe et al. 2020 show that neither shared vocabulary nor joint training is needed (given pre-trained model in language *s*, they froze all parameters except embedding matrix and fine-tuned on new language *t*)
  - Conneau et al. 2020 showed that independent monolingual BERT models can be aligned post-hoc by tying parameters in top layer
  - Dufter et al. 2020 hypothesized that mBERT is multilingual due to limited parameters, forced to share feature across languages

Generalizing Across Languages (2)

- **mBERT** was the **first of many** multilingual transformer-based LMs
- Lample et al. 2019 proposed **XLM** model (stands for cross-lingual LM)
  - **Balanced data for learning tokenizer** (upsampled "small" languages, downsampled "large" languages)
  - Additional objective: **translation language modeling (TLM)**
  - **TLM:** masked language modeling (MLM) but with pairs of parallel sentences (i.e. translations), i.e. alignment signal required
  - Note: **objective no longer self-supervised but supervised**



Generalizing Across Languages (3)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Conneau et al. 2020 proposed XLM-R
  - **Just MLM, no more TLM**
  - Much larger training data
  - Much larger vocabulary: 250K
  - **Significantly outperformed mBERT** on cross-lingual tasks
- Lauscher et al. 2020 showed cross-lingual performance largely dependent on type of "size" of language and "distance" to English
  - **Size:** size of pre-training set
  - **Distance:** how structurally similar it is to English
  - They found performance drops a lot with smaller/more distant languages
- **Number of languages also impactful** to cross-lingual performance
  - Conneau et al. 2020 found a trade-off between performance and number of languages for a fixed model size
  - **Curse of multilinguality:** more languages lead to better cross-lingual performance up to a point, after which performance generally degrades

Zero-Shot vs Few-Shot

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Most works** described **so far** focused on zero**-shot transfer**
  - **Assumption:** zero-labeled examples in target language
  - Perhaps more a scientific question rather than a realistic setting
  - E.g. "can we do it in a zero-shot setting?"
  - Why?
- In practice, almost always possible to label a small number of examples in any target language
  - This would mean a **few-shot cross lingual** setting
- Lauscher et al. 2020 proposed **sequential few-shot transfer**
  - **First**, **fine-tune** LM on **specific task** using **"large" language**, e.g. English
  - **Then**, **fine-tune** on **same task** using **"small" target language**, e.g. Javanese
  - PROs: They found massive improvements over zero-shot setting
  - CONs: first step expensive, not cross-lingual signal about task due to separate steps
- Besides such quantitative studies, **qualitative research** was also done

Cross-lingual Transfer in mBERT (1)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Muller et al. 2021 studied the internal mechanism that allow mBERT do performn cross-lingual transfer.
  - **Recall mBERT:** *pre-trained* on multiple languages, *fine-tuned* on some source language, *performed inference* on another target language
- **First**, they replaced (pre-trained) transformer layers one by one with new, randomly initialized layers, compared performance with mBERT
  - Method for **feature attribution**: what is the impact of this feature on the model prediction?
  - By replacing the layer with random weights and checking difference in performance, we can see how important that layer is for the model
- This was done on **three tasks**:
  - Part-of-Speech (POS) tagging: (assign labels to parts of sentences, e.g. verb)
  - Dependency parsing: predicting dependency between parts of sentence
  - Named Entity Recognition (NER): identify entities in text, e.g. Jane Austen
- Each task evaluated in **two settings**: monolingual, cross-lingual (2 lang.)

Cross-lingual Transfer in mBERT (2)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- They found **lower layers more important in cross-lingual setting**
  - Replacing lower layers with random weights had much more impact cross-lingually
  - Replacing upper layers had little impact in both settings
- This suggests lower layers act as "multilingual encoder", and upper layers as task-specific, language-agnostic "predictor"
  - Upper layers not important for cross-lingual transfer
  - Lower layers much more Important cross-lingually than monolingually

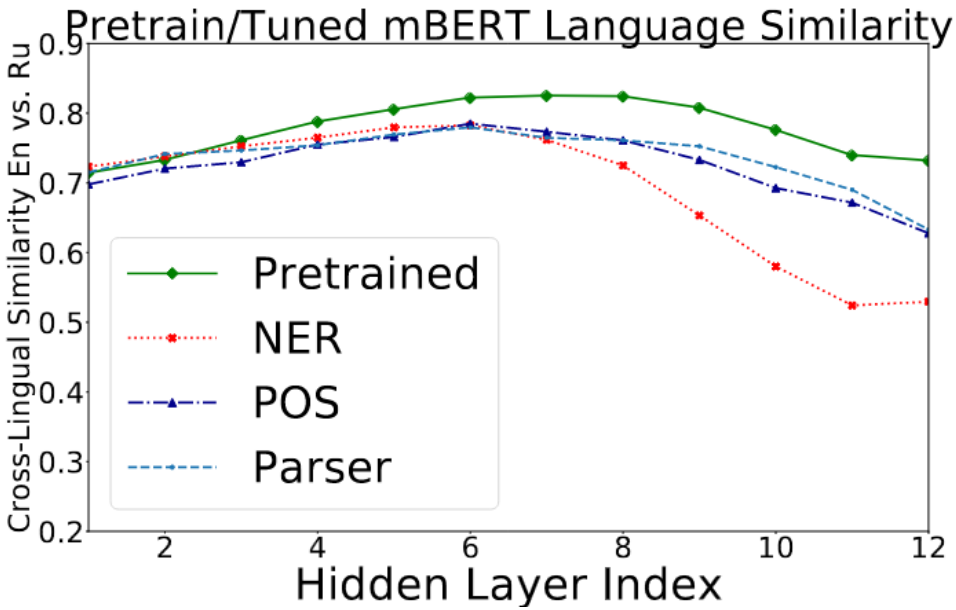
		RANDOM-INIT *of layers*					
SRC-TRG	REF	Δ 1-2	Δ 3-4	Δ 5-6	Δ 7-8	Δ 9-10	Δ 11-12
		Parsing					
EN - EN	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
RU - RU	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
AR - AR	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
EN - X	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - X	55.41	-7.69	-3.71	-3.13	-1.70	0.92	0.94
AR - X	27.97	-4.91	-3.17	-1.48	-1.68	-0.36	-0.14
		POS					
EN - EN	96.51	-0.30	-0.25	-0.40	-0.00	0.05	0.02
RU - RU	96.90	-0.52	-0.55	-0.40	-0.07	0.02	-0.03
AR - AR	79.28	-0.35	-0.49	-0.36	-0.19	-0.05	-0.00
EN - X	79.37	-8.94	-2.49	-1.66	-0.88	0.20	-0.14
RU - X	79.25	-10.08	-2.83	-1.65	-2.74	0.01	-0.45
AR - X	64.81	-6.73	-3.50	-1.63	-1.56	-0.73	-1.29
		NER					
EN - EN	83.30	-2.66	-2.14	-1.43	-0.63	-0.23	-0.12
RU - RU	88.20	-2.08	-2.13	-1.52	-0.64	-0.33	-0.13
AR - AR	87.97	-2.37	-2.11	-0.96	-0.39	-0.15	0.21
EN - X	64.17	-8.28	-5.09	-3.07	-0.79	-0.47	-0.13
RU - X	62.13	-15.85	-9.36	-5.50	-2.44	-1.16	-0.06
AR - X	65.59	-16.10	-8.42	-3.73	-1.40	-0.25	0.67

≥ REF | < REF | ≤ -2 points | ≤ -5 points

Cross-lingual Transfer in mBERT (3)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- They also looked at **alignment per layer**
  - **How?** Average contextualized representations of input sentence in one language, compare with averaged representation in another language



- They found **representations more aligned toward center layers**
  - Suggests model initially learns similar representations across languages
  - May be task dependent, similar results found later (Chi et al., Gaschi et al)

Multilingual Evaluation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **mNLP:** democratization of NLP methods across languages
  - Still an open challenge, e.g. Gemini does not speak javanese
  - Research still ongoing in this direction
- Many **multilingual tasks** proposed for research, e.g.:
  - Americas NLI: natural language inference (NLI) covering 10 low-resource indigenous languages of the Americas
  - MaskhaNER: named-entity recognition (NER) on 10 low-resource african languages
- Today, **common to evaluate LMs on benchmarks with multiple tasks**
  - XGLUE: 11 tasks, 19 languages
  - X-TREME-R: 10 tasks, 50 languages
  - Etc.

Multilinguality in LLMs

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Have LLMs solved mNLP?**
  - In short, **no**.
  - Many languages still not covered
  - Large variance in performance across languages
- Aren't some LLMs multilingual?
  - Yes!
  - Gemini supports over 40 languages
  - Llama-3 supports 30 languages
- But **cross-lingual signals usually not present during pre-training**
  - So, ongoing research to determine whether specific changes to architecture or training regimes are necessary
  - E.g. cross-lingual attention (feed model cross-lingual training examples)
  - Or if simply scaling will get us to the mNLP goal

Summary

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Multilingual NLP:** NLP systems that can process multiple natural languages
- **Cross-lingual transfer:** using knowledge acquired in some source language *s* on a task described in another target language *t*
  - Can be accomplished in many ways depending on task, model architecture
- **Representation alignment:** similar words in different languages clustered together in embedding space
- **Multilingual transformers**: typically pre-trained in multiple languages
  - LMs designed for cross-lingual transfer
  - E.g. via zero-shot cross lingual transfer, sequential cross-lingual fine-tuning
- Modern **LLMs are already multilingual**, they are trained in multiple languages
  - But performance across languages varies a lot given task and language
  - **Open challenge**: improve multilingual performance of LLMs

References

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- References linked in corresponding slides
- Multilingual NLP by Prof. Goran Glavaš