COMP4650/6490 Document Analysis

Multilingual and Low Resource NLP

ANU School of Computing

Administrative matters

- Practice exercise
- Quiz 3
  - Will open later this week
- Final exam:
  - Exam instructions will be available in Week 12
- Drop-in sessions
  - 1pm - 2pm Friday 20 October
  - 1pm - 3pm Thursday 26 October
  - 1pm - 2pm Friday 27 October
  - Location: Room 3.41, Level 3, Hanna Neumann Building

Outline

- **Multilingual NLP**
  - Multilingual word embeddings

- **Low Resource (Language) NLP**
  - Challenges
  - Cross-lingual language model (XLM)

Outline

- **Multilingual NLP**
  - Multilingual word embeddings
- **Low Resource (Language) NLP**
  - Challenges
  - Cross-lingual language model (XLM)

Multilingual NLP

Multilingual NLP deals with cross-lingual resources and models

- There is a need for the development of Multilingual applications, for example, commercial applications including: forum moderation and product recommendation systems
- It's possible to build multilingual NLP resources with monolingual or parallel corpora.... → lots of data

Aims:

- Solve NLP problems which require us to understand multiple languages
- Streamline our NLP tools to work seamlessly across many languages
- Solve NLP problems in languages without a lot of data

Multilingual NLP

Examples of Multilingual NLP problems:

Machine translation

- Translate a sentence from one language into another.
- Typically modelled as a sequence-to-sequence problem.
- Often we use parallel corpora to train the model.

Intra-word code switching

Where the author switches languages part way through a word.

For example, in **oetverkocht** ‘sold out’, the particle **oet** ‘out’ is used that is associated with Limburgish whereas **verkocht** ‘sold’ is associated with Dutch. (Nguyen & Cornips, 2016)

Can be formulated as a sequence labelling problem at the character level, e.g. Tags with BIO encoding

Multilingual resources

Universal Dependencies: POS-Tags, morphological features, syntactic dependencies, and treebanks across 100+ languages

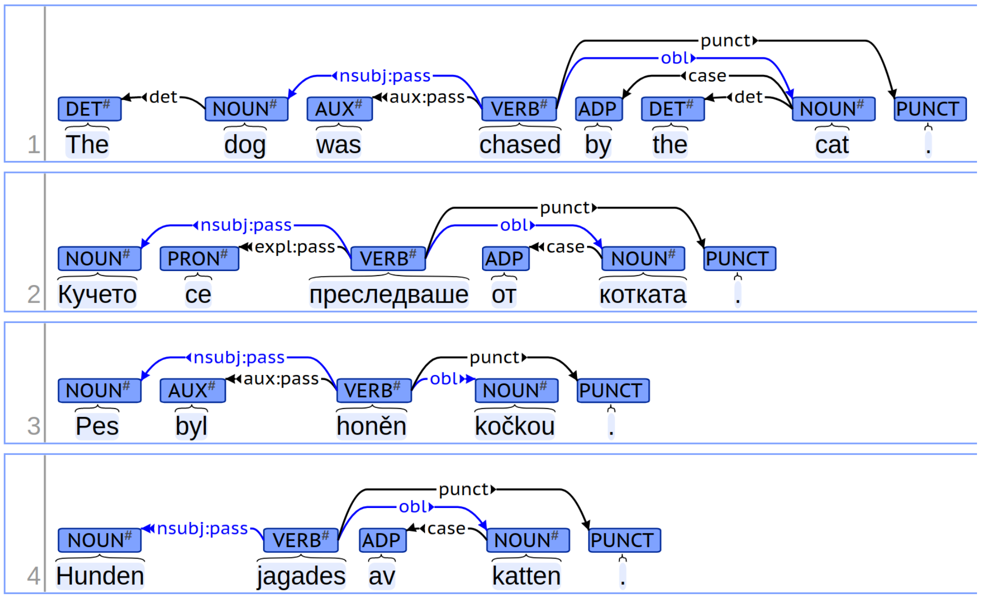
- Facilitates multilingual parser development
- Cross-lingual learning
- Parsing research from a language typology perspective

English:

Bulgarian:

Czech:

Swedish:



https://universaldependencies.org/

The same sentence in four languages parsed using Universal Dependencies.

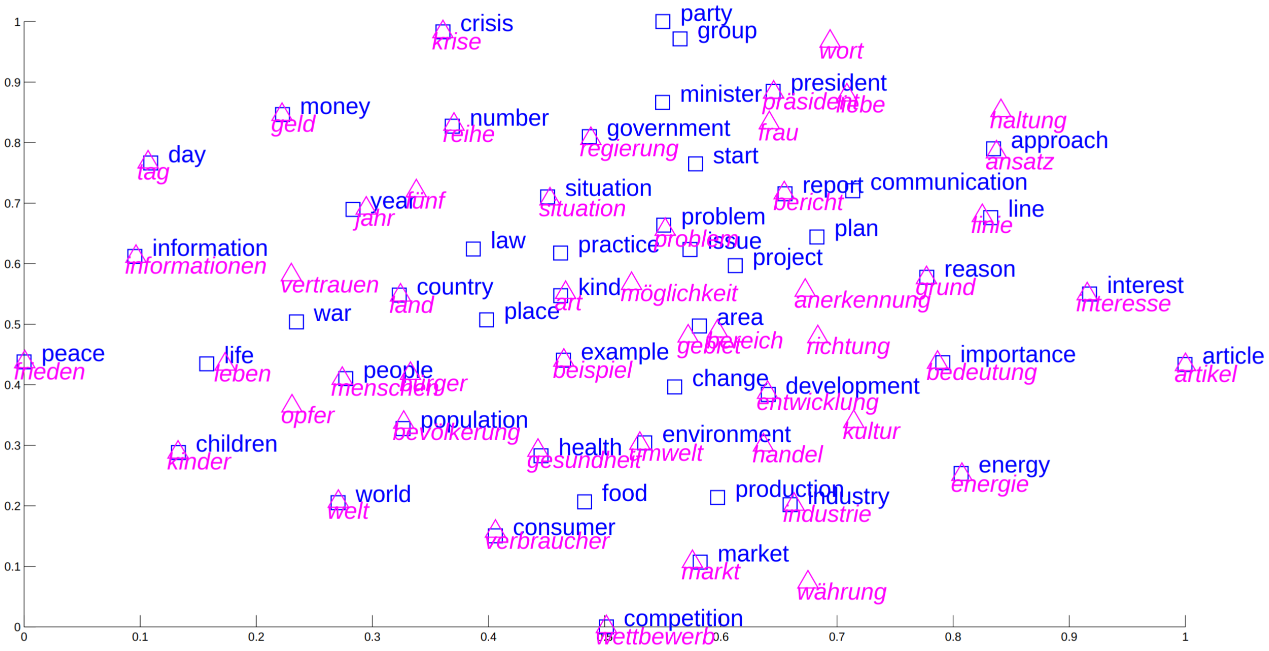
Note:

- the main grammatical relations are the same
- Also, most of same POS tags are used

Multilingual NLP

Multilingual word embeddings

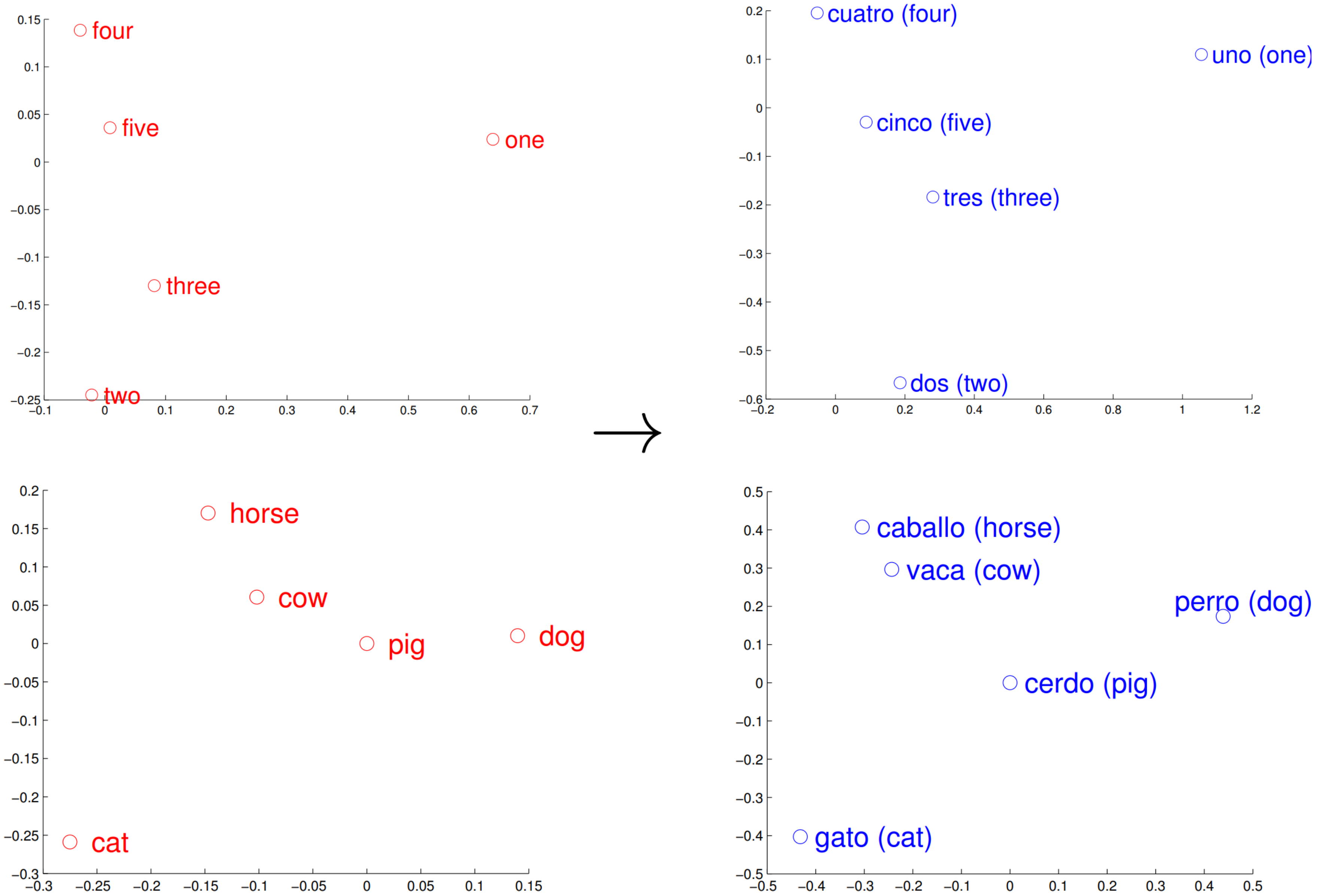
A word embedding space that is shared across multiple languages. Similar words in different languages are close in the embedding space.



(Above) A shared embedding space between English and German (Luong et al., 2015) Constructed using a parallel corpus.

More recent approaches work by aligning mono-lingual embeddings, because it is easier to find large amounts of data.

Multilingual NLP



Similar geometric relations between numbers and animals in English and Spanish
https://arxiv.org/abs/1309.4168

Multilingual NLP

There are several method for constructing multilingual word embeddings:

Supervised: using a small bilingual dictionary to learn a mapping from the source to the target space with (iterative) Procrustes alignment

Unsupervised: without any parallel data, learn a mapping from the source to the target space using adversarial training and (iterative) Procrustes refinement

Supervised approach

Input:

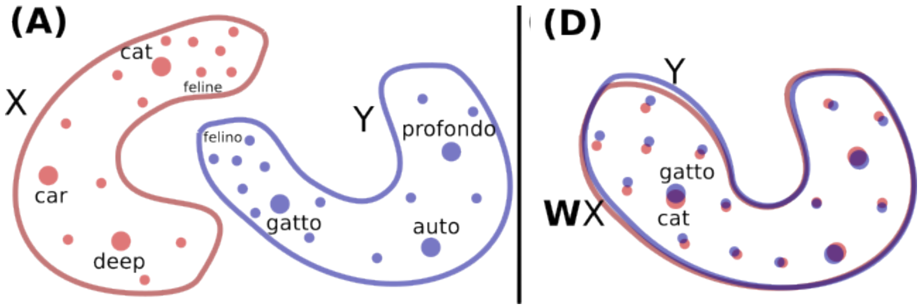
- Word vectors from language A (denoted $x_i \in \mathbb{R}^d$)
- Word vectors from language B (denoted $y_i \in \mathbb{R}^d$)
- A small set of words in language A that are translations of words in language B Represented as word pairs, denoted: $(x_i,\ y_i)_{i \in \{1,\ldots,n\}}$

Output:

- A mapping from vectors in language A to vectors in language B: $y_i \approx Wx_i$

$$X = \begin{bmatrix} 0.12 & 0.4 & \ldots \\ 0.3 & 0.7 & \ldots \\ 0.2 & 0.99 & \ldots \\ \cdots & \cdots & \cdots \end{bmatrix} \quad Y = \begin{bmatrix} 0.32 & 0.21 & \ldots \\ 0.1 & 0.85 & \ldots \\ 0.5 & 0.3 & \ldots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

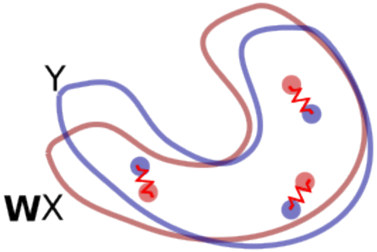
Learn a $d \times d$ transformation matrix W which does the mapping from X to Y.



Supervised approach

We use the word pairs, denoted $(x_i, y_i)_{i \in \{1,\ldots,n\}}$ as supervised examples to optimise a loss function that achieves: $y_i \approx Wx_i$

Minimize: $\frac{1}{n}\sum_{i=1}^{n} \ell(Wx_i, y_i)$



By choosing W, where W is constrained to be an orthogonal mapping (i.e. $W^\top W = I$) (Preserving similarities, e.g. $y_i^\top y_j \approx (Wx_i)^\top (Wx_j) = x_i^\top W^\top W x_j = x_i^\top x_j$)

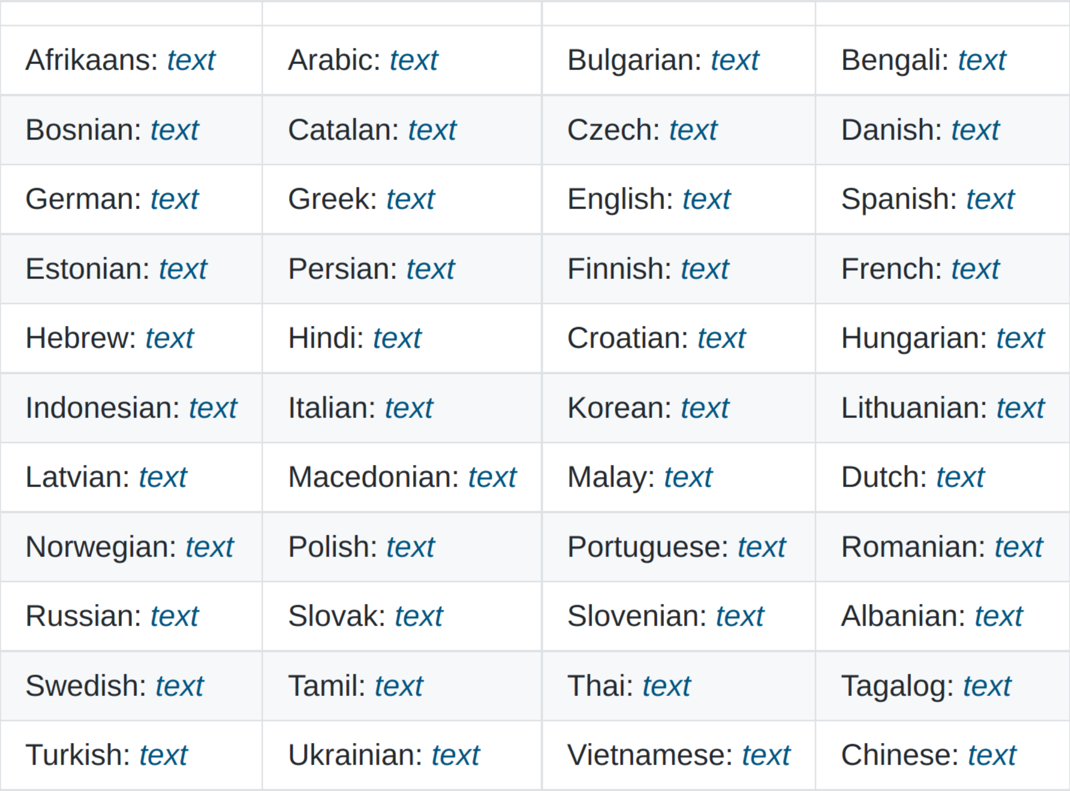
One choice of ℓ is Euclidean distance:

$$\ell(Wx_i, y_i) = \|Wx_i - y_i\|^2$$

More recent techniques have proposed loss functions that work better such as Cross-domain Similarity Local Scaling (CSLS)

Multilingual NLP

Supervised word embeddings for 44 languages, aligned in a single vector space from fastText
https://arxiv.org/abs/1804.07745

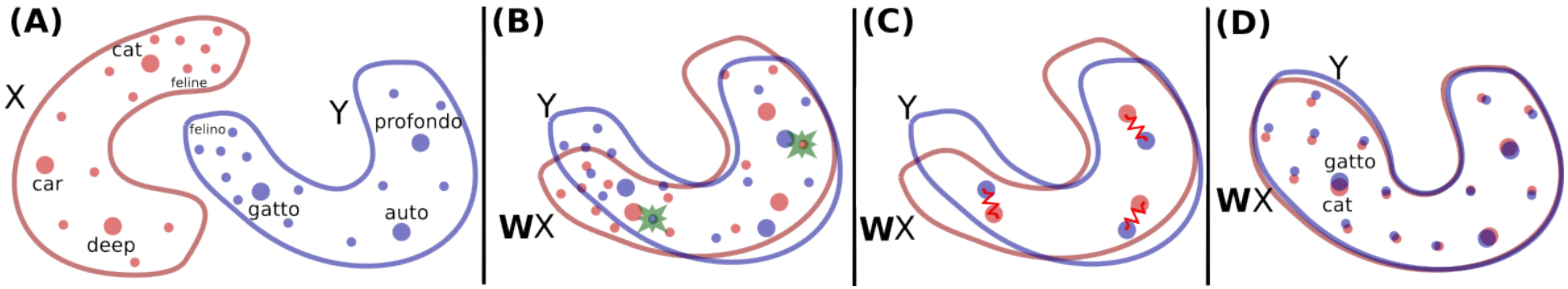


Afrikaans: *text*	Arabic: *text*	Bulgarian: *text*	Bengali: *text*
Bosnian: *text*	Catalan: *text*	Czech: *text*	Danish: *text*
German: *text*	Greek: *text*	English: *text*	Spanish: *text*
Estonian: *text*	Persian: *text*	Finnish: *text*	French: *text*
Hebrew: *text*	Hindi: *text*	Croatian: *text*	Hungarian: *text*
Indonesian: *text*	Italian: *text*	Korean: *text*	Lithuanian: *text*
Latvian: *text*	Macedonian: *text*	Malay: *text*	Dutch: *text*
Norwegian: *text*	Polish: *text*	Portuguese: *text*	Romanian: *text*
Russian: *text*	Slovak: *text*	Slovenian: *text*	Albanian: *text*
Swedish: *text*	Tamil: *text*	Thai: *text*	Tagalog: *text*
Turkish: *text*	Ukrainian: *text*	Vietnamese: *text*	Chinese: *text*

Uses a supervised approach

- Relies on a small number of known word mappings between language pairs.
- Learns a transformation matrix W that maps one language space into another
- Minimises CSLS between transformed words for known word pairs

Multilingual NLP



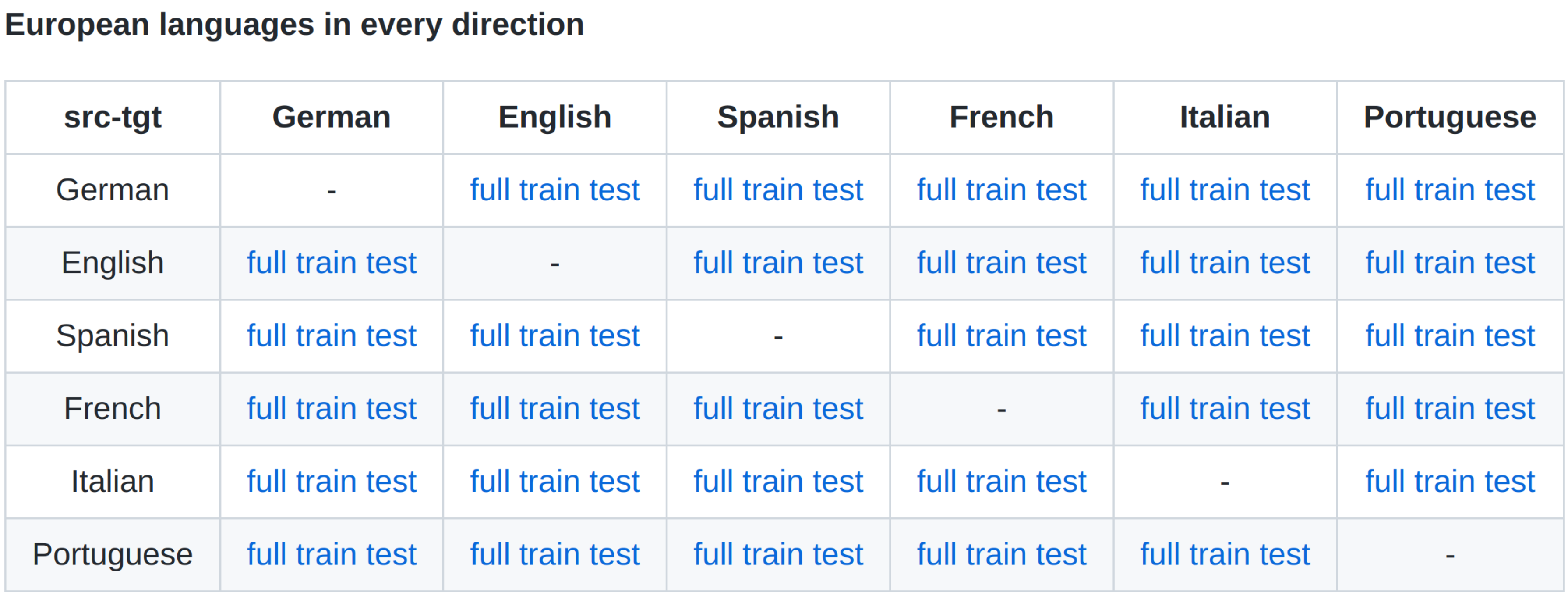
MUSE is a Python library for multilingual word embeddings, whose goal is to provide the community with:

- state-of-the-art multilingual word embeddings (fastText embeddings aligned in a common space)
- large-scale high-quality bilingual dictionaries for training and evaluation

Based on (Conneau et al. 2018) https://arxiv.org/abs/1710.04087

- Unsupervised alignment approach
- Uses a discriminator objective (like a GAN) to get a very good initial alignment
- Refines the alignment using most frequent words in the initial alignment as seeds

Ground-truth bilingual dictionaries from Facebook research for evaluating fastText (and similar methods)



European languages in every direction

src-tgt	German	English	Spanish	French	Italian	Portuguese
German	-	full train test	full train test	full train test	full train test	full train test
English	full train test	-	full train test	full train test	full train test	full train test
Spanish	full train test	full train test	-	full train test	full train test	full train test
French	full train test	full train test	full train test	-	full train test	full train test
Italian	full train test	full train test	full train test	full train test	-	full train test
Portuguese	full train test	full train test	full train test	full train test	full train test	-

Outline

- **Multilingual NLP**
  - Multilingual word embeddings

- **Low Resource (Language) NLP**
  - Challenges
  - Cross-lingual language model (XLM)

Low Resource NLP

Domains where there is a small amount of data

- Applies in domains which don't have substantial training resources: biomedical, legal, etc.

Today we will mostly be talking about low resource languages:

- Languages lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications

- State-of-the-art NLP models require large amounts of training data and complex language-specific engineering

- Language-specific engineering is expensive, requires linguistically trained speakers of the language

Recent survey paper (Hedderich et al, 2021) https://arxiv.org/abs/2010.12309

Why it matters?

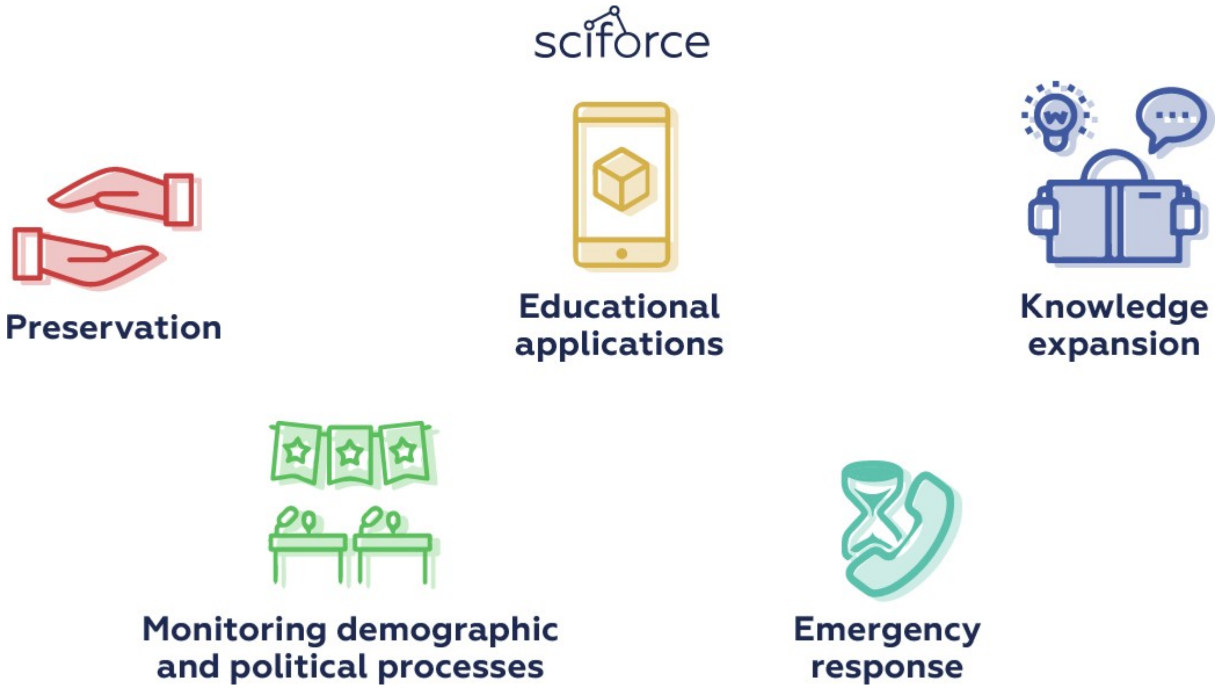
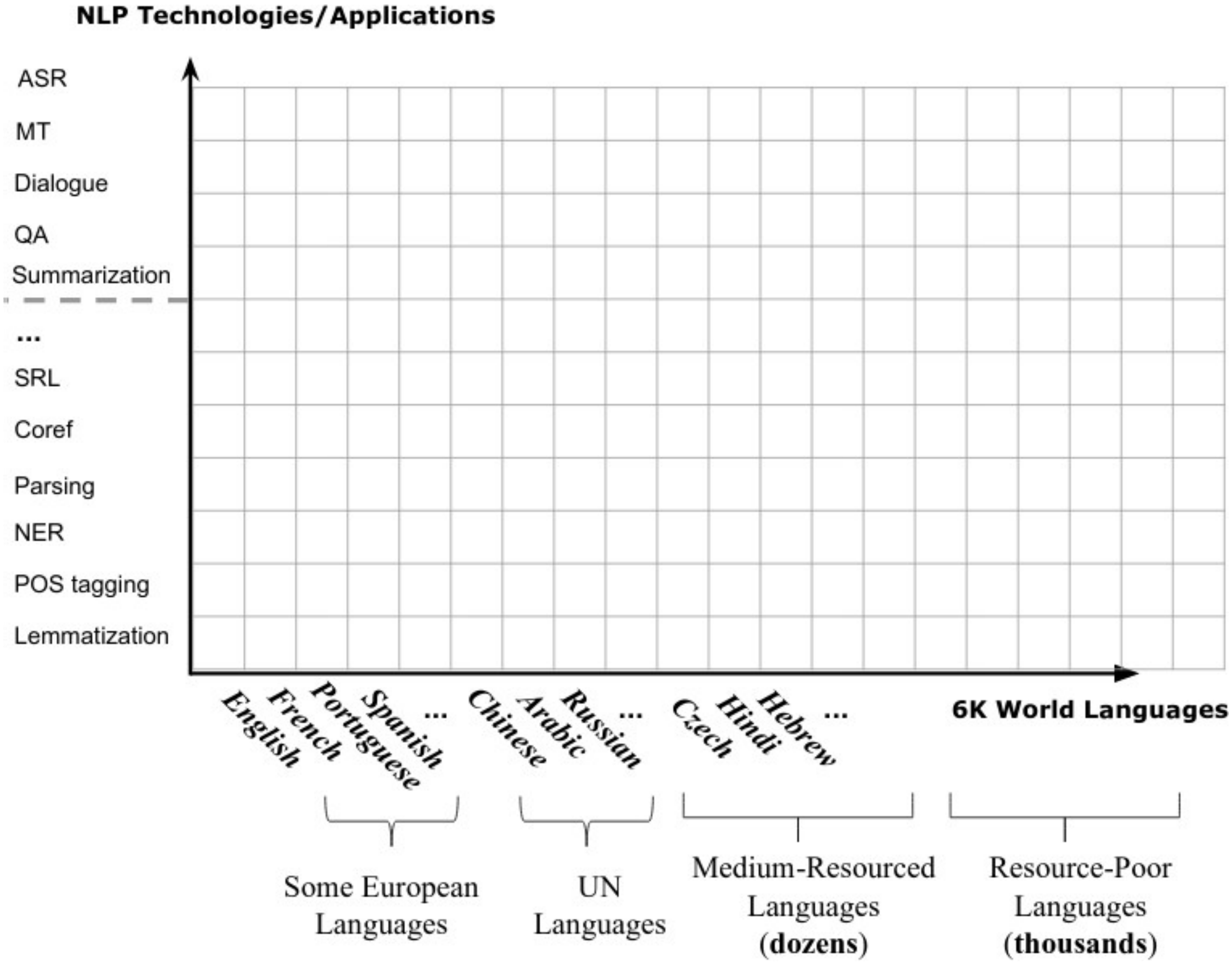


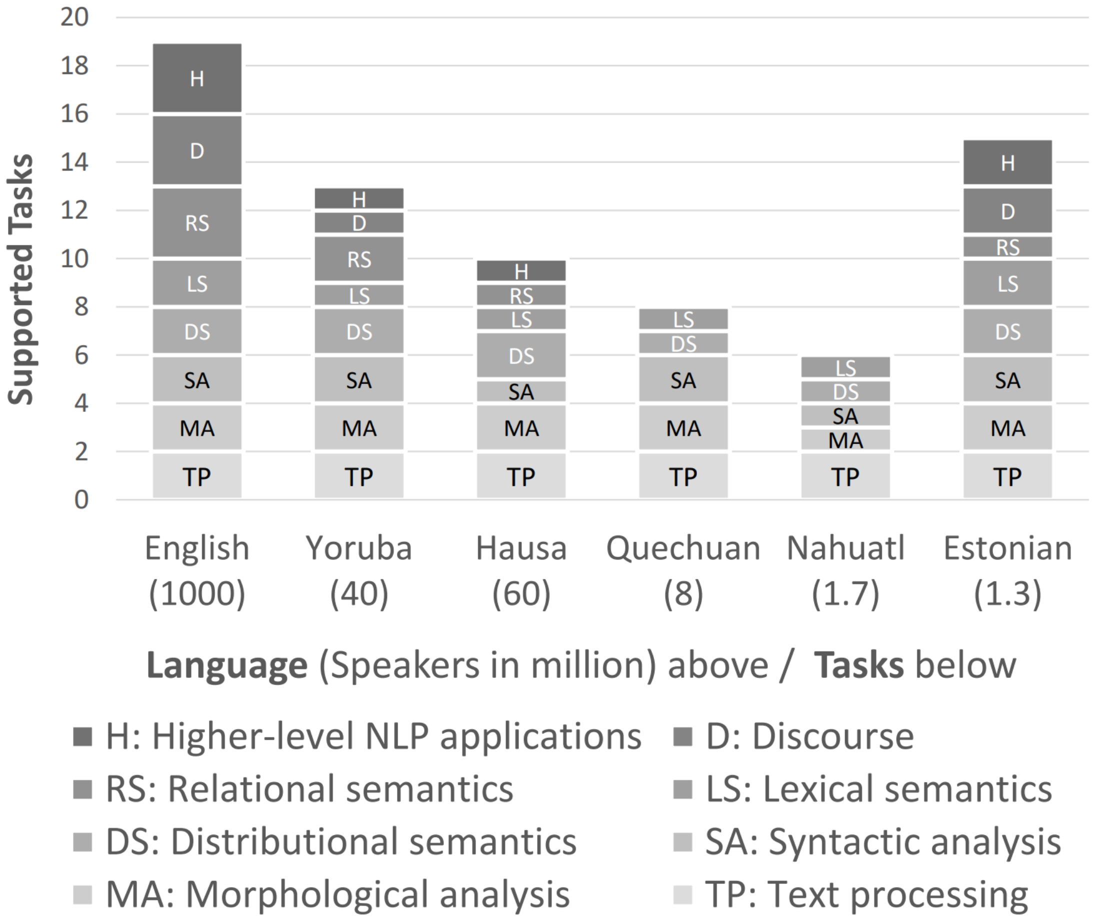
Image source: https://medium.com/sciforce/nlp-for-low-resource-settings-52e199779a79

Low Resource NLP

Low resource does not refer only to minority or endangered languages, but also to languages spoken by millions in developing countries



Low Resource NLP



Languages spoken by millions of people do not have training data for many tasks. (Hedderich et al, 2021)

Australian National University

Low Resource NLP

Examples of machine learning methods suitable for low resource NLP

- Active learning
- Transfer learning
- Multi-task learning
- Learning-to-Learn and Meta-Learning
- Semi-supervised learning
- Dual learning (English → French, French → English)
- Unsupervised learning

Low Resource NLP

Unsupervised Learning:

- Unsupervised POS-Tagging
- Unsupervised dependency parsing
- Brown clustering

Also:

- Universal representations and interlinguas

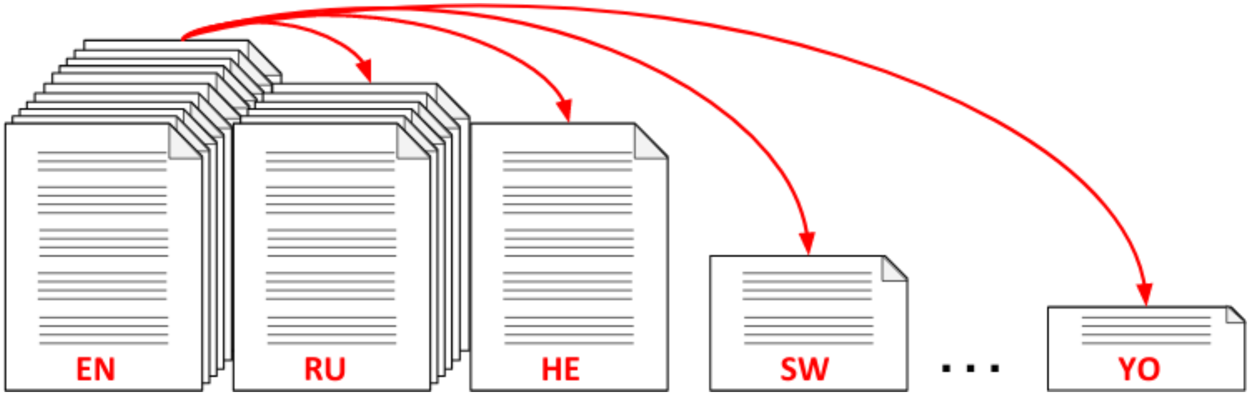
Low Resource NLP

Cross-Lingual Transfer Learning

Transfer of annotations
Such as POS tags, syntactic or semantic features via cross-lingual bridges (e.g. word or phrase alignments)

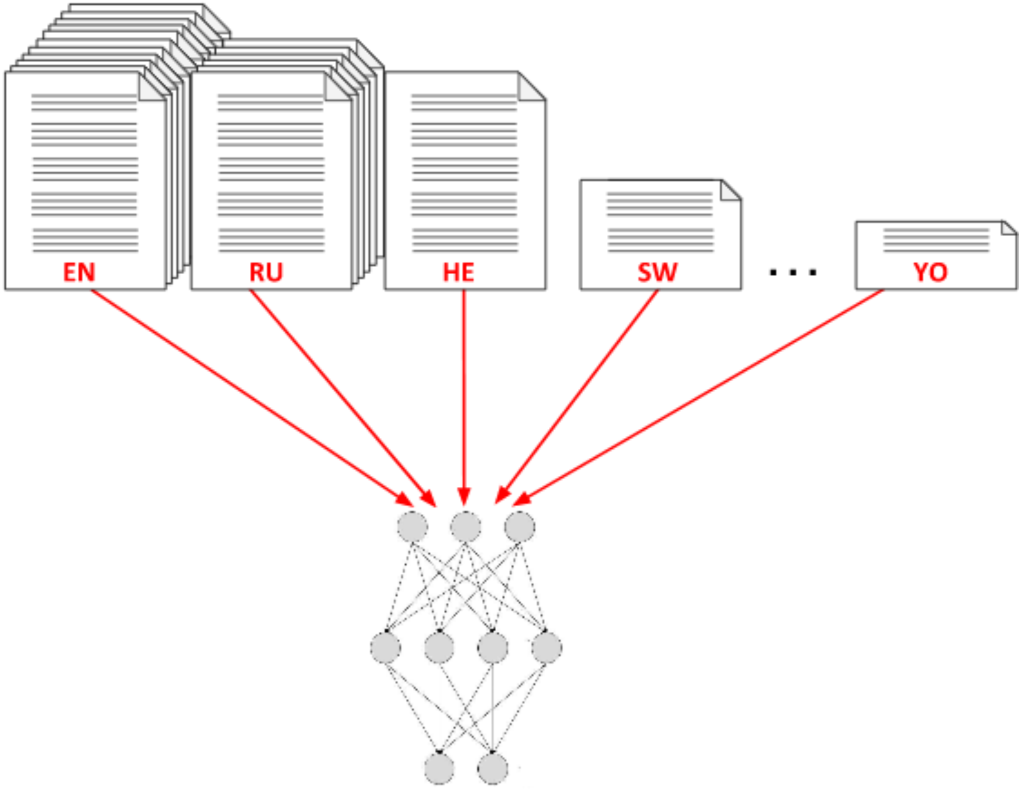
Transfer of models
Training a model in a resource-rich language and applying it in a resource-poor language in zero-shot or one-shot learning

Transfer other parameters, e.g. features



Low Resource NLP

Joint Multilingual or “Polyglot” Learning



Resource-rich and resource-poor learning using a language-universal representation

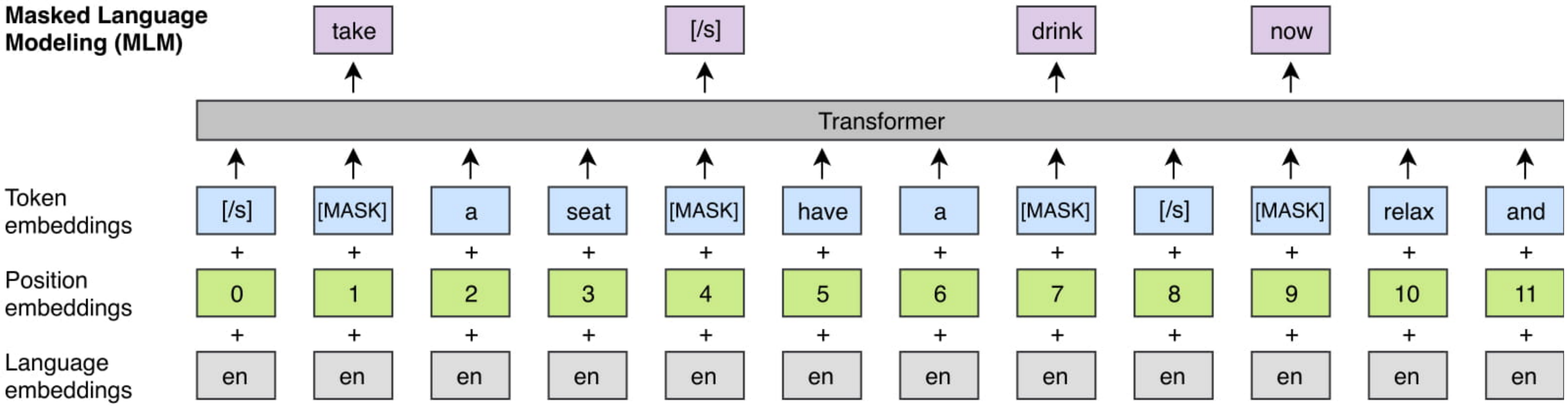
- Convert data in all languages to a shared representation (e.g. multilingual word vectors)
- Train a single model on a mix of datasets in all languages, to enable parameter sharing where possible

Low Resource NLP

XLM is a BERT type model trained on 15 languages:

Trained a single model on 15 languages using Wikipedia text from each language

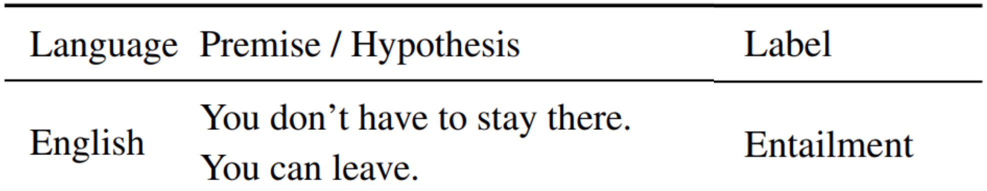
- Sampled sequences from each language during training.
- Used masked language modelling (aka BERT word masking)
- Sub-word vocabulary (Byte Pair Encoding) shared across all languages



The resulting contextual embedding are useful for many different cross language tasks. (Lample & Conneau, 2019)
https://arxiv.org/abs/1901.07291
https://github.com/facebookresearch/XLM

Zero-shot cross-lingual classification

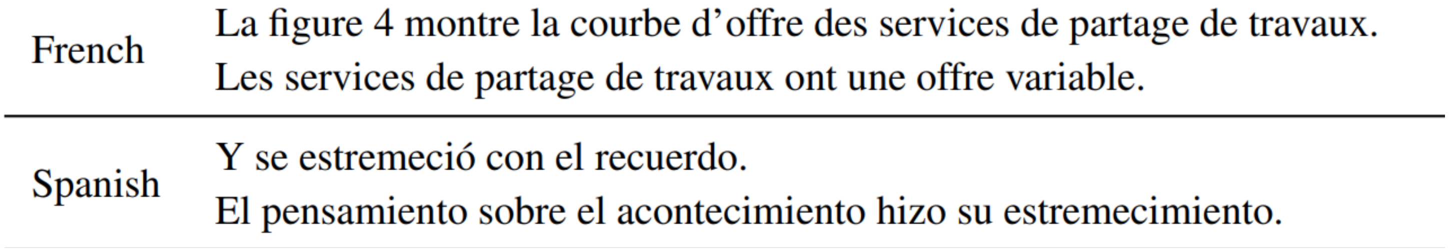
Fine tune XLM on an English language classification task (e.g. NLI with only English text)



Language	Premise / Hypothesis	Label
English	You don’t have to stay there. You can leave.	Entailment

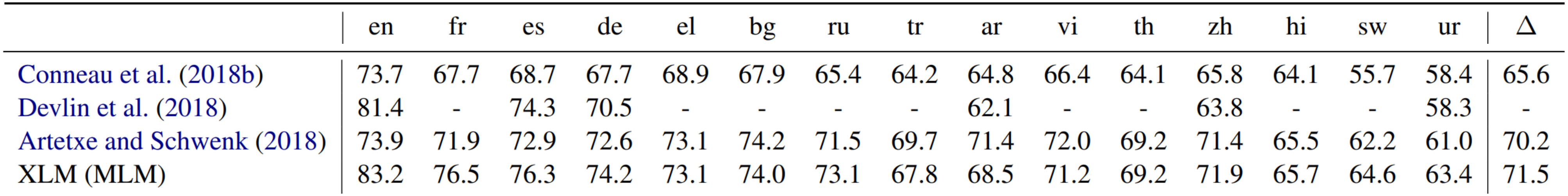
NLI classes are:
Entailment, Contradiction, Neutral

Evaluate on a different language (e.g. French, Spanish ….)



French	La figure 4 montre la courbe d’offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.

Zero-shot cross-lingual classification



	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5

- Conneau et al. (2018b) uses MUSE word embeddings
- Delvin et al. (2018) mBERT is BERT trained on 100 languages but is not quite as good in the above table (perhaps because of the number of languages used) https://github.com/google-research/bert/blob/master/multilingual.md
- Artetxe & Schwenk. (2018) use parallel corpora to train a BiLSTM encoder-decoder
- Newer model: XLM-R (Conneau et al, 2020)
  https://arxiv.org/abs/1911.02116
  Based on a better BERT implementation (RoBERTa)

Summary

- **Multilingual NLP**
  - Multilingual word embeddings
- **Low Resource (Language) NLP**
  - Challenges
  - Cross-lingual language model (XLM)

Australian National University

References

Luong, M.T., Pham, H. and Manning, C.D., 2015, June. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 151-159).

Mikolov, T., Le, Q.V. and Sutskever, I., 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H. and Grave, E., 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745*.

Conneau, A., Lample, G., Ranzato, M.A., Denoyer, L. and Jégou, H., 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Hedderich, M.A., Lange, L., Adel, H., Strötgen, J. and Klakow, D., 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Lample, G. and Conneau, A., 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.