How to evaluate alignment of contextualized embeddings

Félix Gaschi

LORIA (UMR 7503) - Posos company

About me

- 3rd year PhD student at ORPAILLEUR (LORIA)
  - supervisors: Yannick Toussaint, Parisa Rastin
  - topic: Multilingual Information Extraction: with (hopefully) applications to biomedical question answering
  - More broadly working with multilingual language models
- CIFRE thesis: funded by Posos, Paris
  - Securing prescription with a platform for physicians
  - ≈40 people
  - 5 members of the R&D team
  - Since 2017

OUR GOAL

Context

Multilingual contextualized embeddings can **solve some cross-lingual tasks** but there is **no consensus on the alignment** of representations.

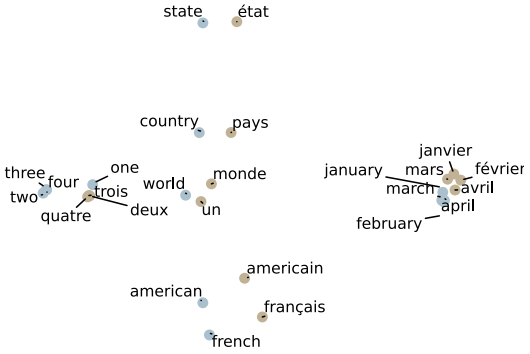
Question

Are mBERT contextualized representations of words from different languages aligned?

Outline

- Introduction on multilingual contextualized embeddings
- How to evaluate alignment in multilingual Transformers
- Proposed method using bilingual dictionaries

MULTILINGUAL WORD EMBEDDING: ALIGNING MONOLINGUAL WORD2VEC



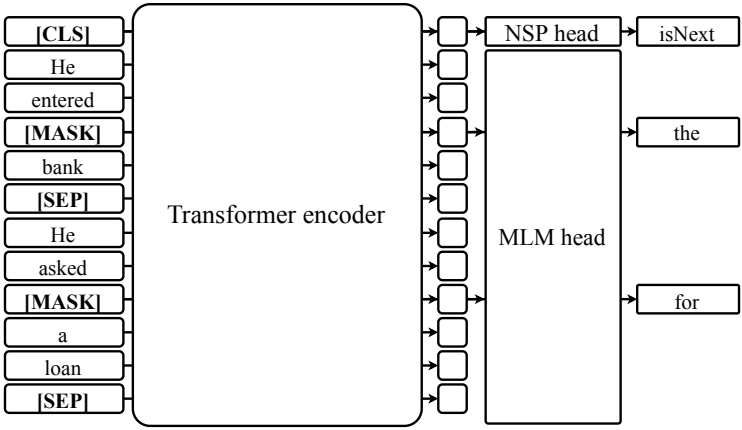
Søgaard et al., "Cross-Lingual Word Embeddings"

Typical strategy for learning aligned static word embedding

1. learn monolingual embeddings (word2vec, FastText...)
2. **explicitely** learn a mapping

our baseline: "aligned FastText"

CONTEXTUALIZED EMBEDDINGS: BERT



Output representation depends on the context.

Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

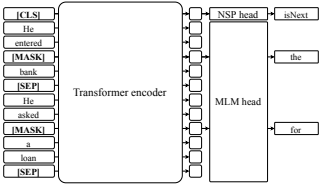
CONTEXTUALIZED EMBEDDINGS ARE USEFUL FOR TRANSFER LEARNING

Static word embeddings can be used as pretrained input of larger models.

Contextualized embedding are already large models. They can be fine-tuned on a downstream task (NER, NLI, QA)

- Replace MLM head with classification head (linear classifier)
- Fine-tune (re-train) the model on the specific task

Outperforms models trained ”from scratch”



MULTILINGUAL CONTEXTUALIZED EMBEDDINGS: MBERT

Released with BERT but no dedicated papers.

Training data

The 100 most frequent languages of Wikipedia. **No explicit cross-lingual signal**

Sampling strategy

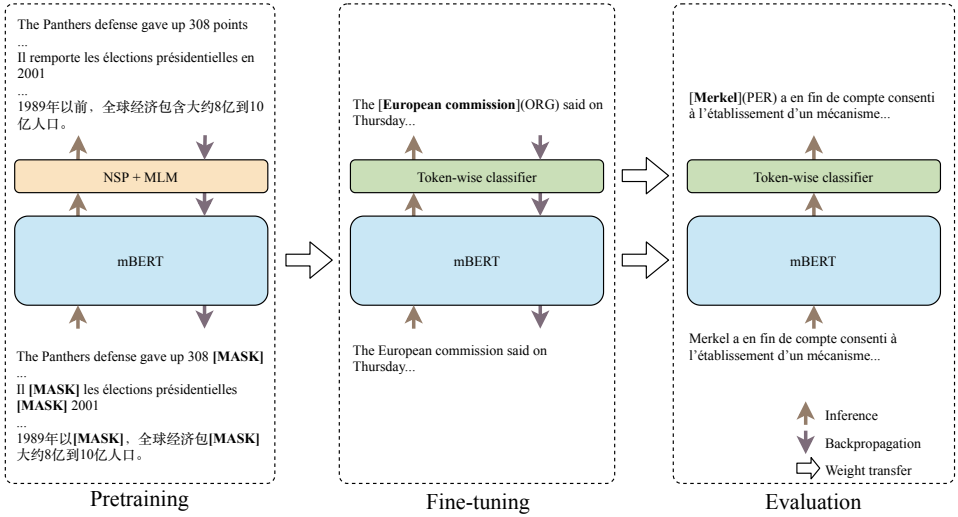
Smoothing of the distribution of languages $p_i = f_i^{0.7}/Z$
English is sampled 100x more often than Icelandic (instead of 1,000)

Vocabulary

Shared and bigger than BERT (110k instead of 30k)
Ideographs are character-tokenized

Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

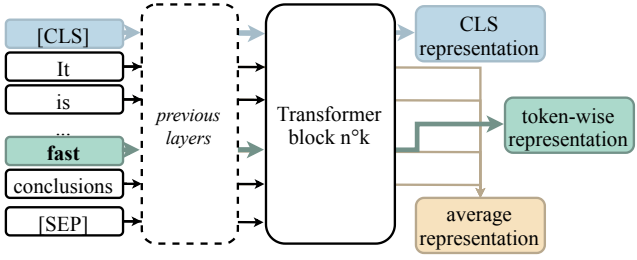
mBERT ALLOWS CROSS-LINGUAL TRANSFER LEARNING



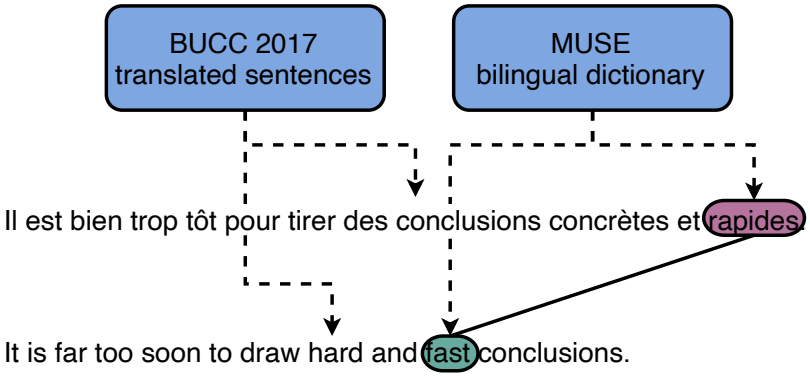
How to evaluate alignment of contextualized embeddings?

Key choices

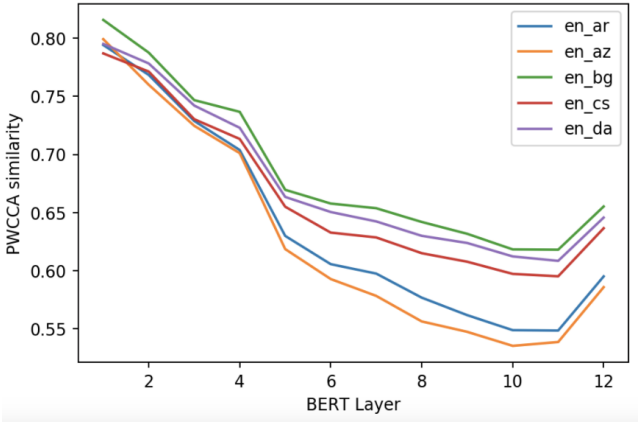
- **What representations?** sentence (CLS or avg), word, subword...
- **What data?** translation dataset, bilingual dictionary, FastAlign...
- **What measure?** distribution of distance, compare with random pairs, NN-search...



WHAT DATA FOR WORD-LEVEL ALIGNMENT



THE SIMILARITY OF THE CLS TOKEN

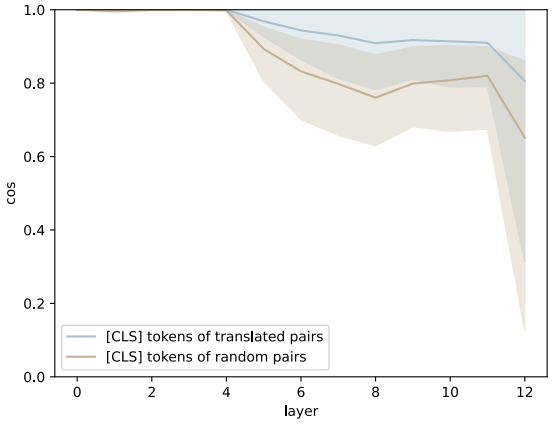


- Representation: CLS
- Data: translation (XNLI)
- measure: PWCCA similarity

Conclusion: ”Bert is not an interlangua”

Singh et al., “BERT is Not an Interlingua and the Bias of Tokenization”

PROBLEM: WE NEED A DISTRACTOR



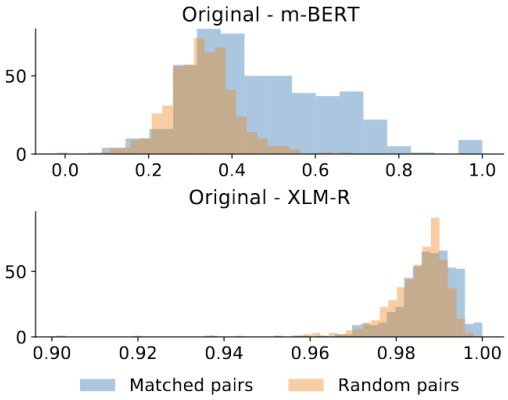
Measuring CLS similarity allows to compare between pairs of languages.

But we need to compare with unrelated pairs of sentences.

Similarity does decrease but actually less than random pairs.

Gaschi et al., Multilingual Transformer Encoders: a Word-Level Task-Agnostic Evaluation

COMPARING SIMILARITY DISTRIBUTIONS



- Representation: token
- Data: translation + FastAlign
- measure: compared cos sim

Conclusion: not aligned (significant overlap)

First issue: FastAlign is error-prone
Second issue: what does the overlap really mean?

Zhao et al., "Inducing Language-Agnostic Multilingual Representations"

OVERLAP $\neq$ BAD ALIGNMENT

What does overlap really mean?

Some unrelated pair is more similar than some related pair:

$$\exists (a, b) \in \text{unrelated}, (c, d) \in \text{related}, d(a, b) < d(c, d) \tag{1}$$

But those two pairs do not necessarily involve the same words

Paired-difference measure

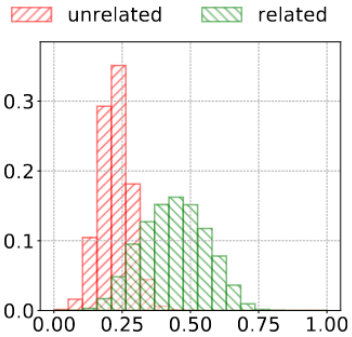
Compare related pairs of words with unrelated pairs involving the same words:

$$D(t_i) = \cos(t_i, s_i^{\text{rel}}) - \frac{1}{n} \sum_{k=1}^{n} \cos(t_i, s_{ik}^{\text{unrel}}) \tag{2}$$

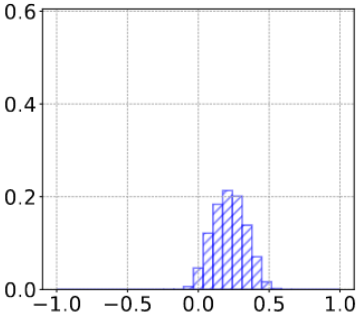
Efimov et al., "The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer"

PAIRED-DIFFERENCE HISTOGRAMS FOR MBERT

Similarity of related and unrelated pairs



Paired-difference histogram



- Representation: token
- Data: translation + FastAlign
- measure: paired difference

Conclusion: well-aligned on the word level

NEAREST-NEIGHBOR SEARCH

With paired-difference, a word is closer to its translation than the **average** unrelated word. But is it closer than most?

NN-search

Inspired from Bilingual dictionary induction for static embeddings.

- Sample N pairs (u_i, v_i)
- NN-search: for each u_i take $j' = \arg\min_j d(u_i, v_j)$
- Measure top-1 accuracy: proportion of cases where $j' = i$

Similar to paired-difference

Take paired-difference

$$D(t_i) = \cos(t_i, s_i^{\text{rel}}) - \frac{1}{n}\sum_{k=1}^{n} \cos(t_i, s_{ik}^{\text{unrel}})$$

NEAREST-NEIGHBOR SEARCH

With paired-difference, a word is closer to its translation than the **average** unrelated word. But is it closer than most?

NN-search

Inspired from Bilingual dictionary induction for static embeddings.

- Sample N pairs (u_i, v_i)
- NN-search: for each u_i take $j' = \arg\min_j d(u_i, v_j)$
- Measure top-1 accuracy: proportion of cases where $j' = i$

Similar to paired-difference

Take all source-words from other pairs for unrelated pairs:

$$D'(u_i) = \cos(u_i, v_i) - \frac{1}{n} \sum_{j=1, j \neq i}^{n} \cos(u_i, v_j)$$

NEAREST-NEIGHBOR SEARCH

With paired-difference, a word is closer to its translation than the **average** unrelated word. But is it closer than most?

NN-search

Inspired from Bilingual dictionary induction for static embeddings.

- Sample N pairs (u_i, v_i)
- NN-search: for each u_i take $j' = \arg\min_j d(u_i, v_j)$
- Measure top-1 accuracy: proportion of cases where $j' = i$

Similar to paired-difference

Replace average with max:

$$D''(u_i) = \cos(u_i, v_i) - \max_{j \neq i} \cos(u_i, v_j)$$

NEAREST-NEIGHBOR SEARCH

With paired-difference, a word is closer to its translation than the **average** unrelated word.
But is it closer than most?

NN-search

Inspired from Bilingual dictionary induction for static embeddings.

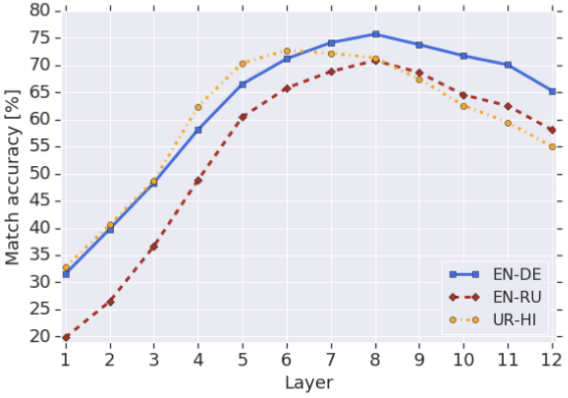
- Sample N pairs (u_i, v_i)
- NN-search: for each u_i take $j' = \arg\min_j d(u_i, v_j)$
- Measure top-1 accuracy: proportion of cases where $j' = i$

Similar to paired-difference

Compute proportion of $D''(u_i) > 0$ rather than distribution:

$$\text{accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[D''(u_i) > 0\right]$$

USING NN-SEARCH FOR SENTENCE REPRESENTATIONS



- Representation: sentence (average), **centered by language**
- Data: translation
- measure: NN-search

Conclusion: aligned... but representations are centered

Pires, Schlinger, and Garrette, "How Multilingual is Multilingual BERT?"

WHY CENTERING REPRESENTATIONS?

Centering representations

For each sampled translation pair (u_i, v_i), we replace it with:

$$(u_i', v_i') = \left(u_i - \frac{1}{n}\sum_{j=1}^{n} u_j, v_i - \frac{1}{n}\sum_{j=1}^{n} v_j \right) \tag{3}$$

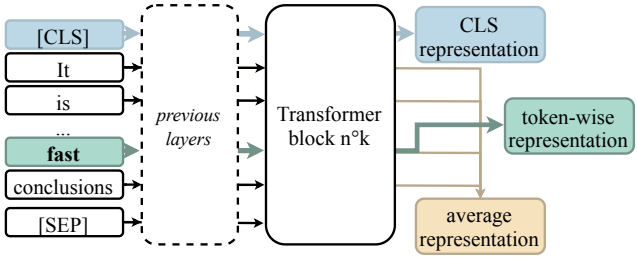
Why centering?

There seem to be a language-specific component in mBERT

- linear classifier predicts the language with ≈95% accuracy
- Centering can improve zero-shot cross-lingual transfer

Pires, Schlinger, and Garrette, "How Multilingual is Multilingual BERT?"
S. Wu and Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT"
Libovický, Rosa, and Fraser, "On the Language Neutrality of Pre-trained Multilingual Representations"

ABSENCE OF CONSENSUS ON THE ALIGNMENT



- There might be a good word-level alignment
  - Similarity of related and unrelated words overlap a lot. But overlap $\neq$ bad alignment
  - Paired-difference: related words are closer than the average unrelated words. But is this enough?
- Sentence-level alignment is good, but after centering, using average representation

OUR PROPOSED METHOD

A word-level, dictionary-based, nearest-neighbor search

- What representations?
  → **Words**
- What data? translation dataset with related word extraction
  → **Rely on bilingual dictionaries** instead of FastAlign
- What measure?
  → **Perform a nearest-neighbor search** among sampled pairs

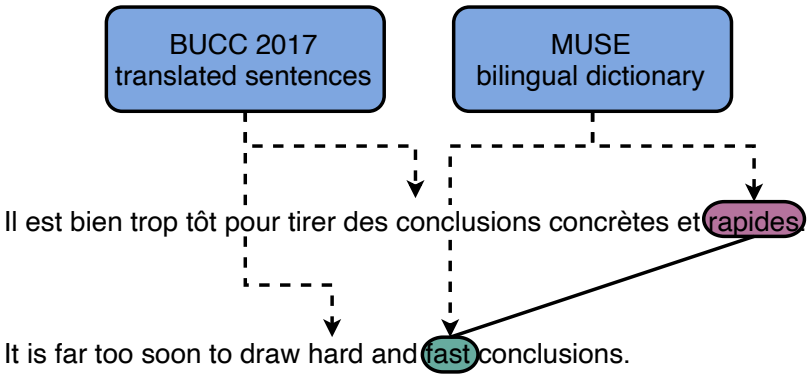
OUR PROPOSED METHOD

A word-level, dictionary-based, nearest-neighbor search

- What representations?
  → **Words**
- What data? translation dataset with related word extraction
  → **Rely on bilingual dictionaries** instead of FastAlign
- What measure?
  → **Perform a nearest-neighbor search** among sampled pairs
- Is this enough?
  → No, **evaluate weak and strong alignment**

Our proposed method

1. Find contextualized translated pairs of words

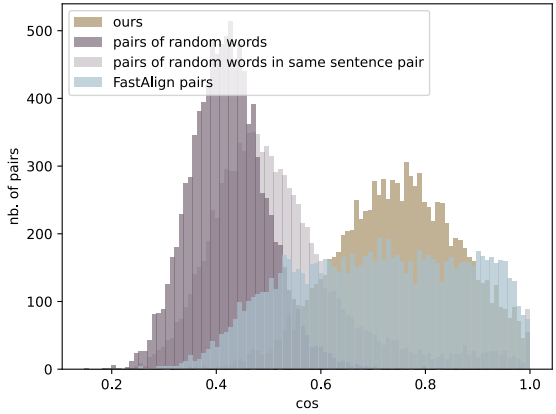


WHY DO WE USE A BILINGUAL DICTIONARY?

Alternative: use probabilistic alignment tool like FastAlign
But FastAlign introduces errors in the extracted pairs

Table: Precision of the extracted pairs

method	en-de	en-fr	ro-en
dictionary	**90.1**	**95.2**	**94.5**
FastAlign	71.3	80.0	71.8



(en-zh)

Dyer, Chahuneau, and Smith, "A Simple, Fast, and Effective Reparameterization of IBM Model 2"

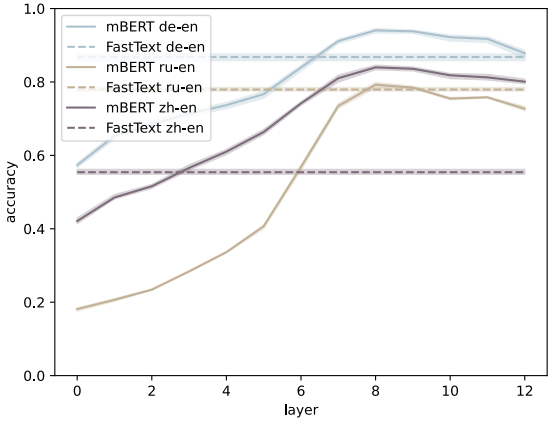
Our proposed method

2. Perform nearest-neighbor search

NN-search

- Extract pairs of contextualized words
- For comparison with static embeddings, keep one occurence for each word
- Sample N=10k pairs (u_i, v_i)
- NN-search: for each u_i take $j' = \arg\min_j d(u_i, v_j)$
- Measure top-1 accuracy: proportion of cases where $j' = i$

RESULTS OF OUR EVALUATION METHOD



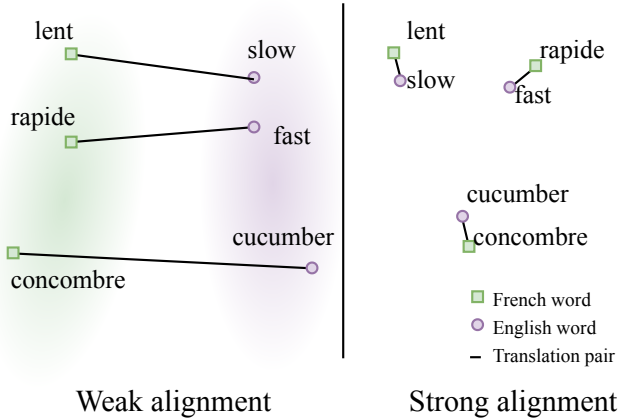
For deeper layers, mBERT is comparable to aligned FastText.

Table: Evaluation for the last layer

model	de-en	ru-en	zh-en
FastText	86.8	77.9	55.4
mBERT	**87.9**	**72.7**	**80.1**
XLM-R	77.4	53.5	54.4
XLM-15	28.5	4.6	11.2

more detailed results in the paper

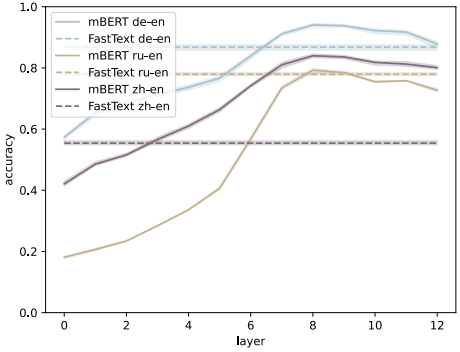
STRONG ALIGNMENT VS. WEAK ALIGNMENT



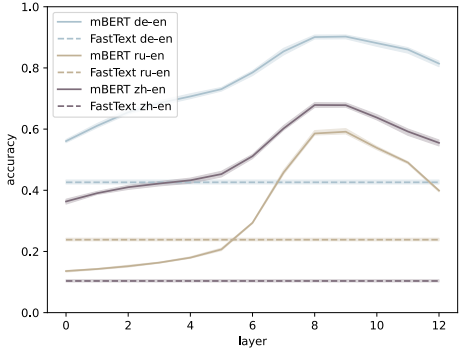
Roy et al., "LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool"

RESULTS FOR STRONG ALIGNMENT

Weak alignment



Strong alignment



mBERT more ”robustly aligned” than FastText aligned.

Conclusion

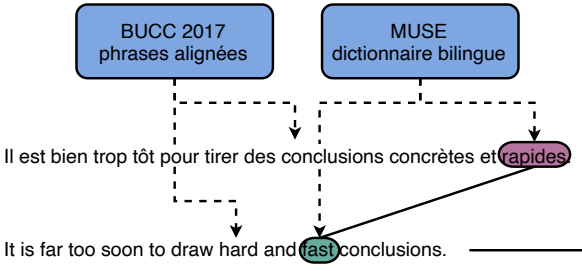
A method for evaluating alignment

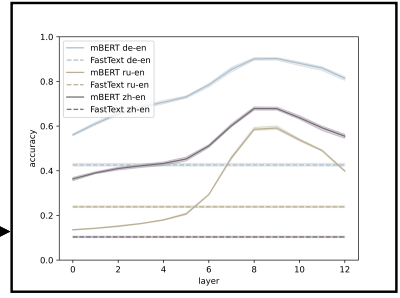
1. Extract contextualized translated pairs with bilingual dictionary
2. Perform Nearest-Neighbor search (using strong alignment)

Word-level alignment of multilingual Transformers

- Deeper layers are better aligned than explicitely aligned baseline
- mBERT has higher scores than others (even those trained with cross-lingual signal)
- Also in the paper: error analysis, importance of context, sentence-level alignment

THANK YOU !





felix.gaschi@loria.fr

BIBLIOGRAPHY I

Bojanowski, Piotr et al. “Enriching Word Vectors with Subword Information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. ISSN: 2307-387X.

Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

BIBLIOGRAPHY II

- Dyer, Chris, Victor Chahuneau, and Noah A. Smith. "A Simple, Fast, and Effective Reparameterization of IBM Model 2". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 644–648. URL: https://aclanthology.org/N13-1073.
- Efimov, Pavel et al. "The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer". In: *ArXiv* abs/2204.06457 (2022).
- Gaschi, Félix et al. *Multilingual Transformer Encoders: a Word-Level Task-Agnostic Evaluation*. 2022. DOI: 10.48550/ARXIV.2207.09076. URL: https://arxiv.org/abs/2207.09076.
- Joulin, Armand et al. "Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

BIBLIOGRAPHY III

Libovický, Jindřich, Rudolf Rosa, and Alexander Fraser. "On the Language Neutrality of Pre-trained Multilingual Representations". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1663–1674. DOI: 10.18653/v1/2020.findings-emnlp.150. URL: https://aclanthology.org/2020.findings-emnlp.150.

Pires, Telmo, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://www.aclweb.org/anthology/P19-1493.

Bibliography IV


Roy, Uma et al. “LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5919–5930. DOI: 10.18653/v1/2020.emnlp-main.477. URL: https://aclanthology.org/2020.emnlp-main.477.


Singh, Jasdeep et al. “BERT is Not an Interlingua and the Bias of Tokenization”. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 47–55. DOI: 10.18653/v1/D19-6106. URL: https://aclanthology.org/D19-6106.


Søgaard, Anders et al. “Cross-Lingual Word Embeddings”. In: *Synthesis Lectures on Human Language Technologies* 12 (June 2019), pp. 1–132. DOI: 10.2200/S00920ED2V01Y2

BIBLIOGRAPHY V

- Vaswani, Ashish et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91Paper.pdf.
- Wu, Shijie and Mark Dredze. “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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077.
- Wu, Yonghui et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

BIBLIOGRAPHY VI

Zhao, Wei et al. “Inducing Language-Agnostic Multilingual Representations”. In: *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Online: Association for Computational Linguistics, Aug. 2021, pp. 229–240. DOI: 10.18653/v1/2021.starsem-1.22. URL: https://aclanthology.org/2021.starsem-1.22.