Crosslingual sentence representations

UNIVERSITY
OF TURKU

Objective

- Word embeddings is a familiar concept by now
- But can it be expanded to larger units of text and maintain the properties seen for word embeddings?
- Can we induce sentence embeddings for sentences like we induced word embeddings for words?

UNIVERSITY OF TURKU

"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

German Kruszewski
Facebook AI Research
germank@fb.com

Guillaume Lample
Facebook AI Research
Sorbonne Universités
glample@fb.com

Loïc Barrault
Université Le Mans
loic.barrault@univ-lemans.fr

Marco Baroni
Facebook AI Research
mbaroni@fb.com

https://yoavartzi.com/sp14/slides/mooney.sp14.pdf

Applications

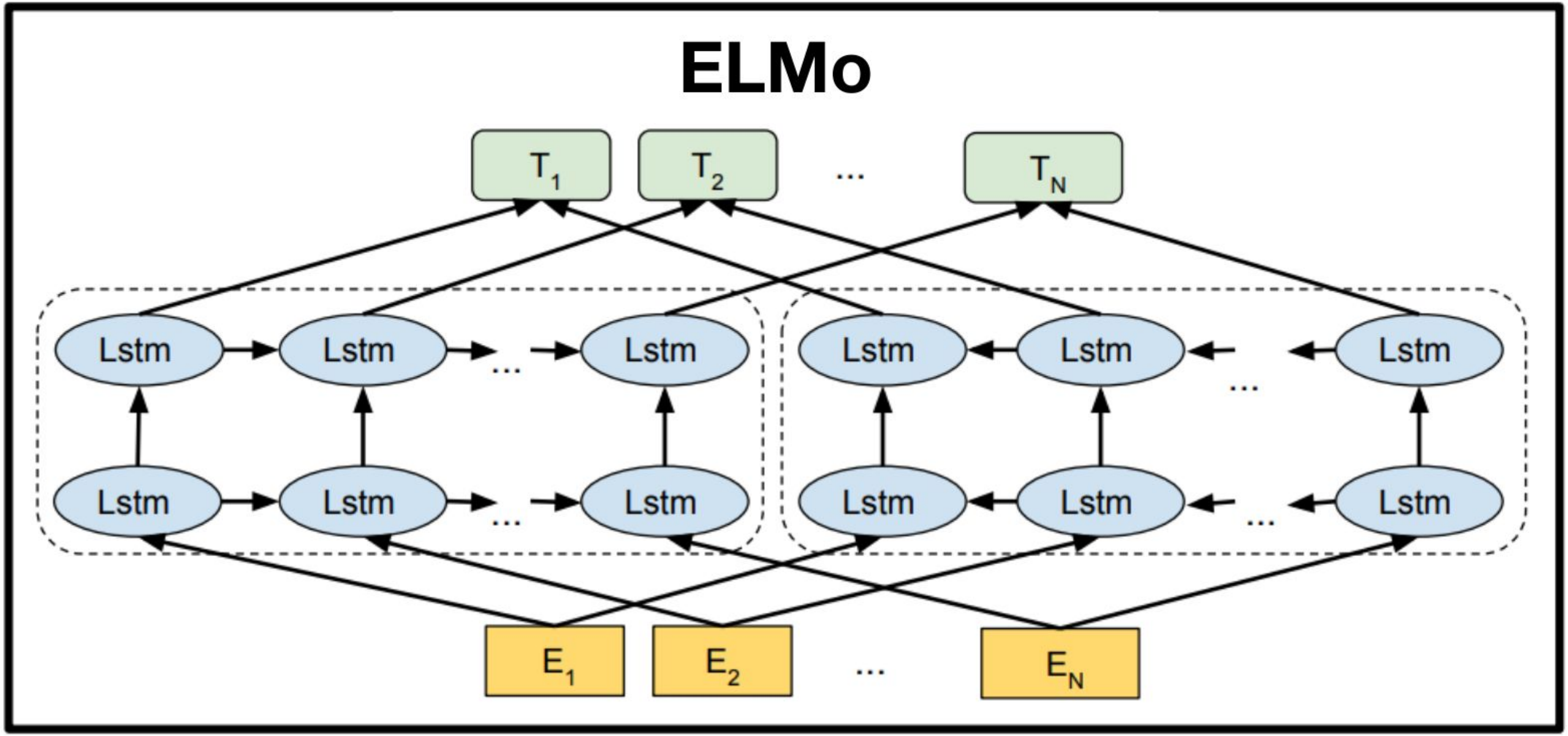
- Fuzzy search / paraphrase detection
- Machine translation
- Document / text classification of any kind
- Transfer models
- Multi- and cross-lingual methods
- Text generation
- ...

Sentence representations

- Monolingual - induced from single language data
- Multilingual - induced from texts of several languages, typically (but not necessarily) translation data

Sentence representations

- You have seen several by now, maybe not thinking about them as such...
  - ELMo: pooled or final LSTM state
  - BERT: [CLS] token representation
- These are used in various classification tasks but not primarily seen as sentence representations in their own right



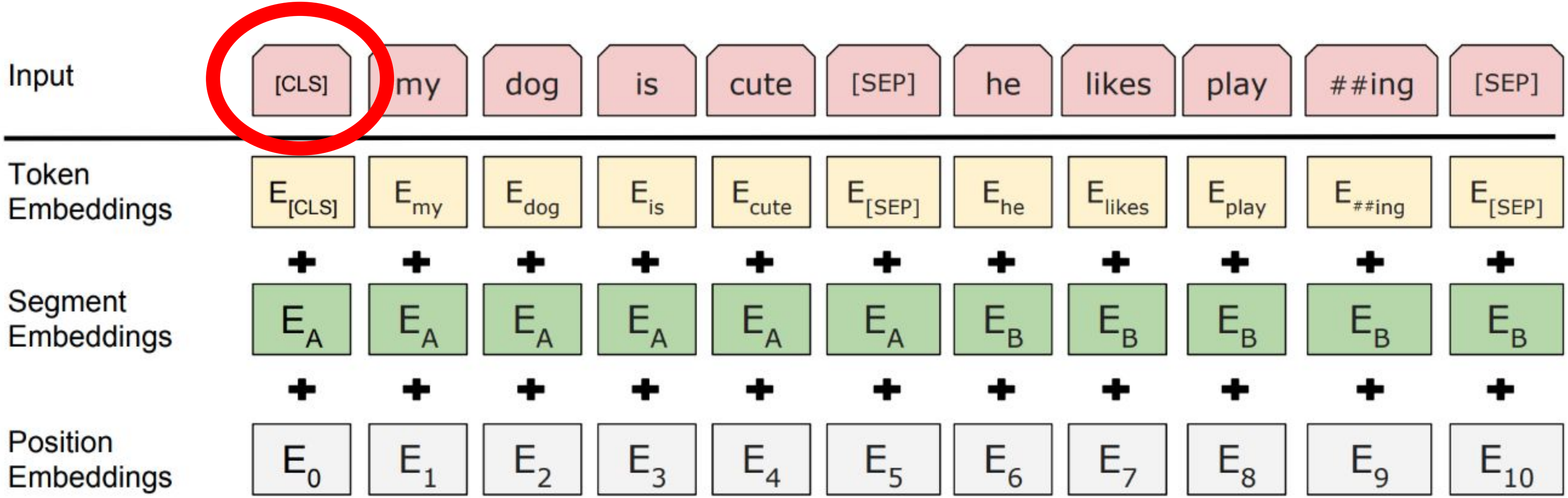
Sentence representation: pooled or final state from ELMo

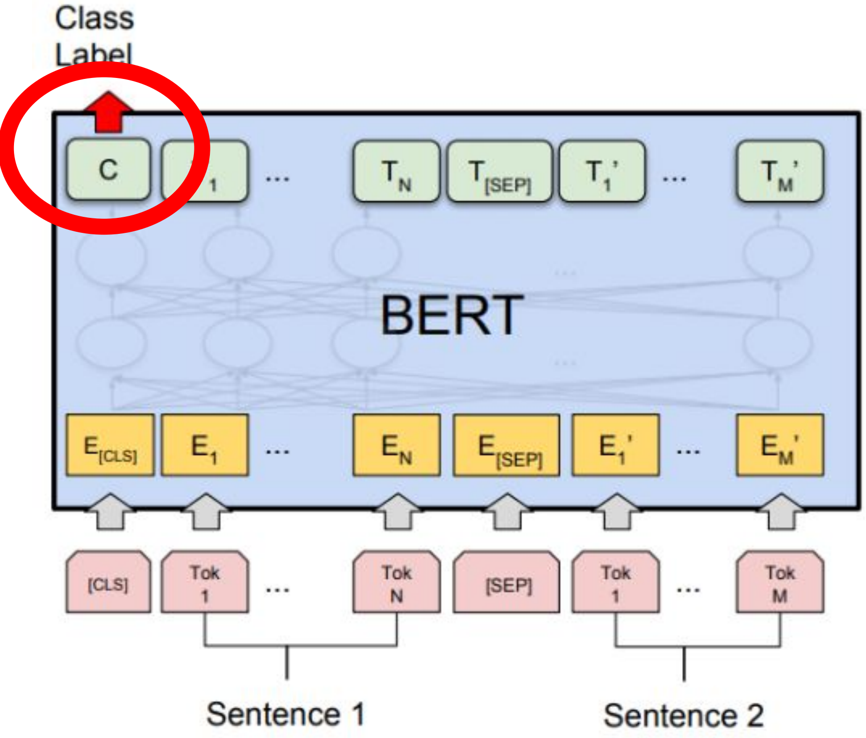
UNIVERSITY OF TURKU

BERT

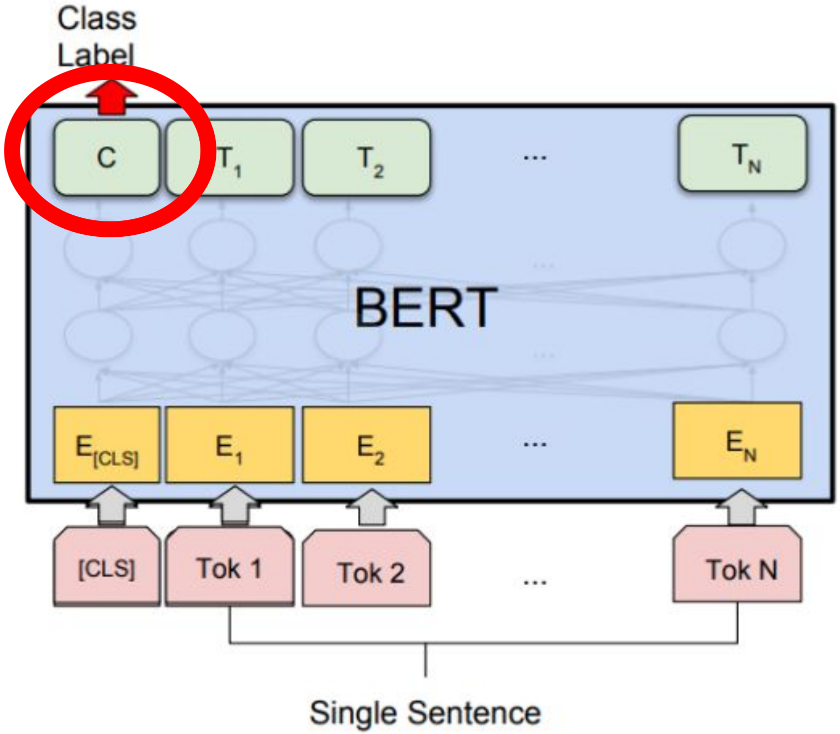
Challenge: learning relationships between sentences

Solution: next sentence prediction





(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA

UNIVERSITY
OF TURKU

Methods for sentence embeddings

- Strive explicitly to produce sentence embeddings as the primary objective, not as a byproduct of some other process
- The surrounding research is interested specifically in the embeddings and their properties

Skip-thought

- Skip-thought (as in skip-gram), a word2vec-like approach for sentences

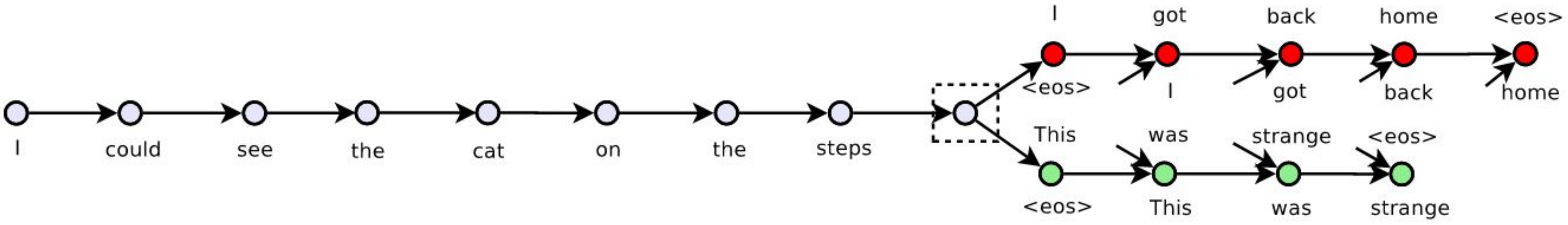


Figure 1: The skip-thoughts model. Given a tuple (s_{i-1}, s_i, s_{i+1}) of contiguous sentences, with s_i the i-th sentence of a book, the sentence s_i is encoded and tries to reconstruct the previous sentence s_{i-1} and next sentence s_{i+1}. In this example, the input is the sentence triplet *I got back home. I could see the cat on the steps. This was strange.* Unattached arrows are connected to the encoder output. Colors indicate which components share parameters. ⟨eos⟩ is the end of sentence token.

https://arxiv.org/pdf/1506.06726.pdf

Query and nearest sentence
he ran his hand inside his coat , double-checking that the unopened letter was still there . he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .
im sure youll have a glamorous evening , she said , giving an exaggerated wink . im really glad you came to the party tonight , he said , turning to her .
although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this . although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .
an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim . a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .
if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa . if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .
then , with a stroke of luck , they saw the pair head together towards the portaloos . then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its flanks .
" i 'll take care of it , " goodman said , taking the phonebook . " i 'll do that , " julia said , coming in .
he finished rolling up scrolls and , placing them to one side , began the more urgent task of finding ale and tankards . he righted the table , set the candle on a piece of broken plate , and reached for his flint , steel , and tinder .

Table 2: In each example, the first sentence is a query while the second sentence is its nearest neighbour. Nearest neighbours were scored by cosine similarity from a random sample of 500,000 sentences from our corpus.

Cross-lingual data

- Translation pairs are a useful source of training data for various sentence embedding induction tasks
- Sentence embedding:
  - A vector encoding the meaning of the sentence
  - ...trying to generate sentence translation from the vector means that the vector should encode the meaning of the sentence well
  - => translation tasks might give raise to useful embeddings!

Encoder

- Input: sequence of word vectors
  - Or sub-words such as BPE / SentencePiece
- Output: **a single vector**
- Architecture - pick your favorite
- Seen in literature:
  - Deep averaging network
  - CNN + max pooling
  - (Bi)LSTM (final state or max pooling)
  - Transformer

Reminder from last week:

Byte-pair encoding (Gage 1994): find most common pair of consecutive bytes in corpus, replace with new byte, repeat

```
onerously dogmatic iconoclast
```

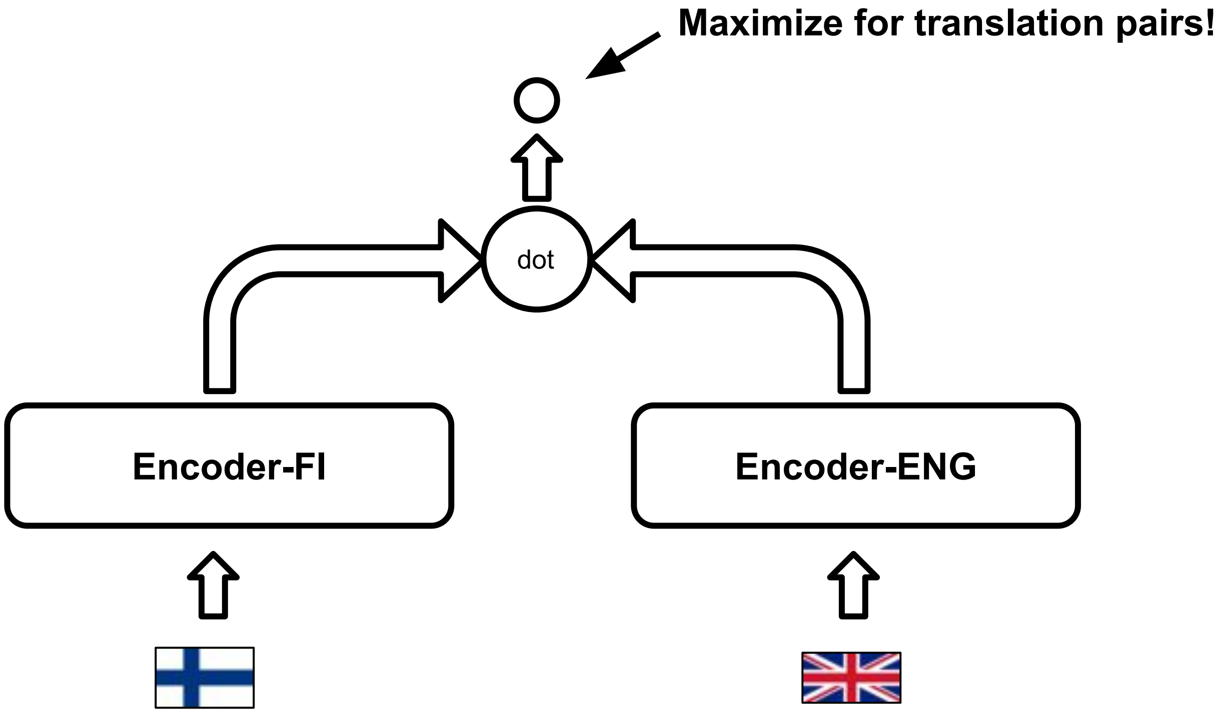
⬇

```
one ##rous ##ly dog ##matic icon ##oc ##last
```

Multilingual embeddings

- Adds a specific requirement of “same sentence in different language gets a similar embedding”
- Similar to multilingual word embeddings where same words in different languages receive similar vectors

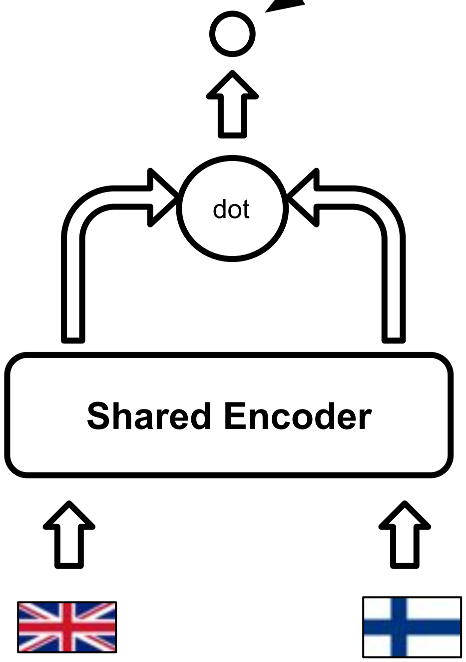
Architecture #1



UNIVERSITY OF TURKU

Architecture #2

Maximize for translation pairs!



UNIVERSITY OF TURKU

Training

- Binary classification problem: translation pair or not?
- Parallel data needed as source of positive pairs
- Negative pairs needed for training as well
- Minimizes distance of positive pairs, maximizes distance of negative pairs

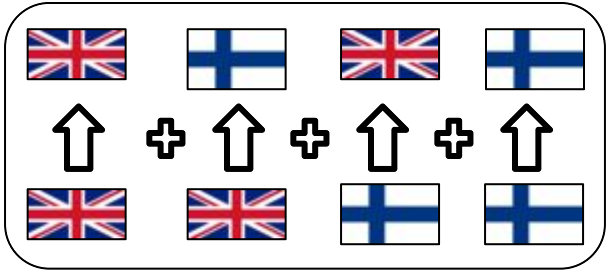
Sampling negatives

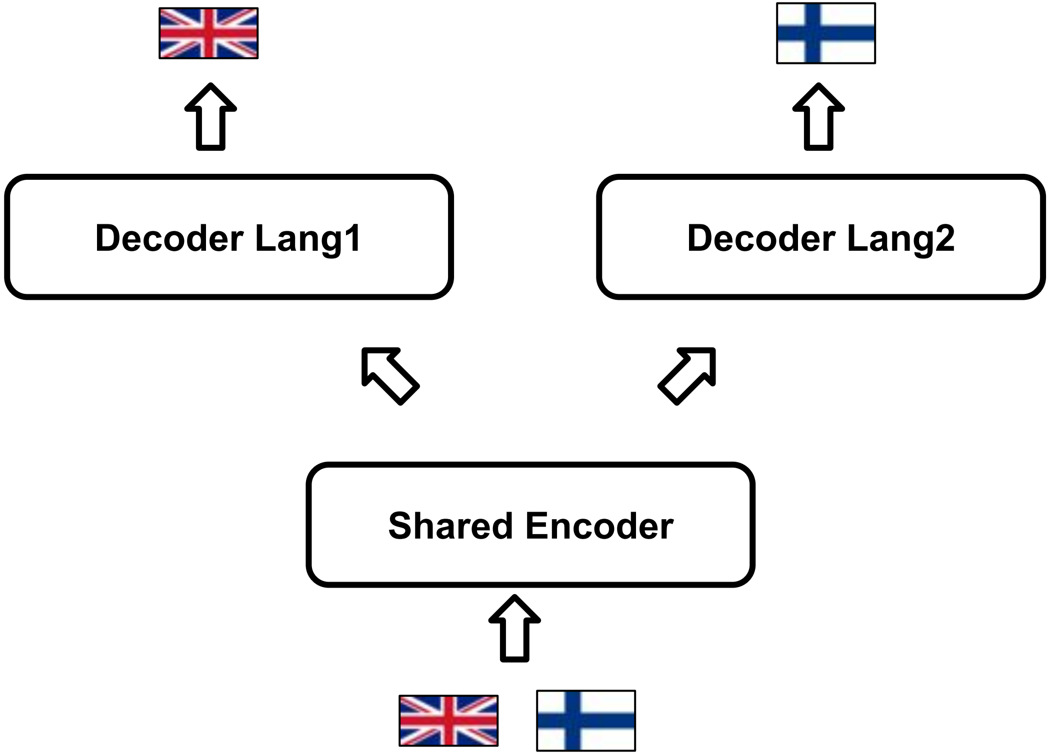
- Choice of negative pairs somewhat problematic
- Random choice
  - Too easy
  - Encoder learns to look for punctuation, personal pronouns, negation..
- Random + length controlled
  - Still too easy
  - Even worse: doesn’t learn to pick same-length sentences

Decoder

- Input: a single vector representing a sentence
- Output: the sentence itself
- Generated character / subword / word at a time
- Architecture - pick your favorite:
  - Left-to-right LSTM
  - Transformer

Architecture #3



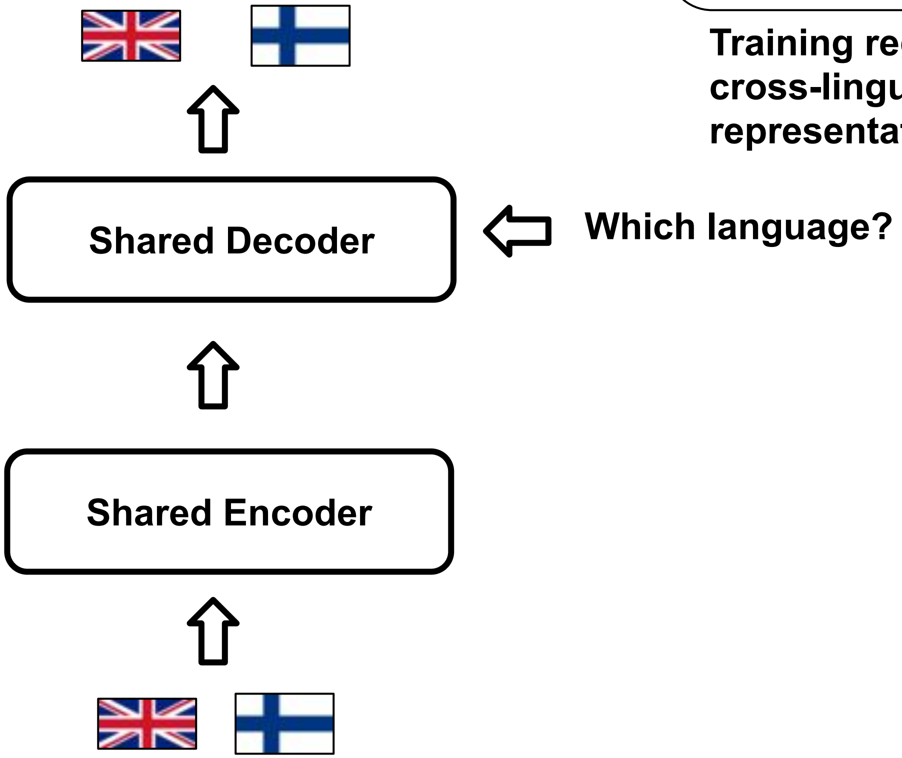


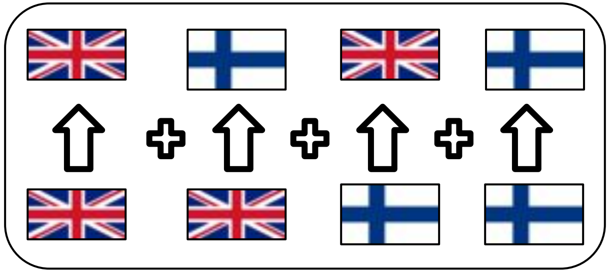
Training regime encourages cross-lingually comparable representations

Constitutes a simple neural machine translation system

UNIVERSITY OF TURKU

Architecture #4





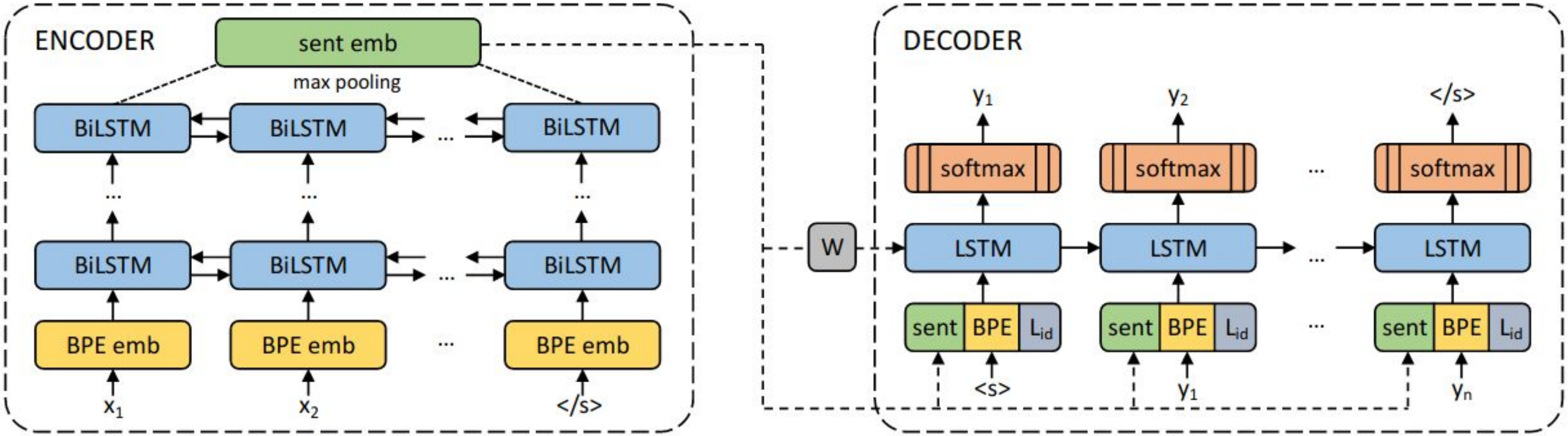
Training regime encourages cross-lingually comparable representations

UNIVERSITY OF TURKU

Training

- Parallel data needed
- No negative samples necessary
  - The decoder enforces the encoder learning meaningful representations
- After training, discard the decoder, keep the encoder

FB LASER



https://arxiv.org/pdf/1812.10464.pdf
https://research.fb.com/downloads/laser-language-agnostic-sentence-representations/