Learning Bilingual Sentence Embeddings via Autoencoding and Computing Similarities with a Multilayer Perceptron

Yunsu Kim, Hendrik Rosendahl*, Nick Rossenbach, Jan Rosendahl, Shahram Khadivi*, Hermann Ney

*eBay Inc., Aachen, Germany

Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany

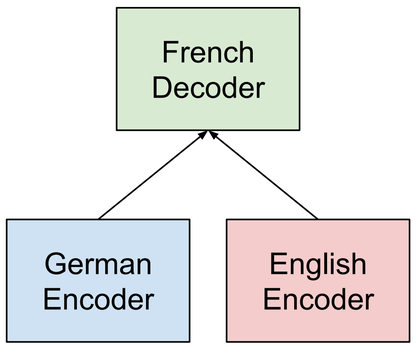
Cross-lingual Sentence Embeddings via NMT

Problem: Get cross-lingual sentence embeddings using bilingual corpora

Solution: NMT with multiple encoders and a single **shared** decoder

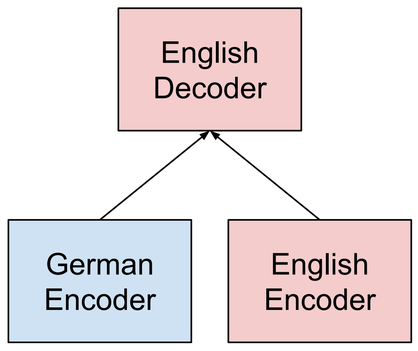
- Encoders should provide consistent representations to the decoder
- Example: cross-lingual sentence embedding for `de/en`

Pivot-based



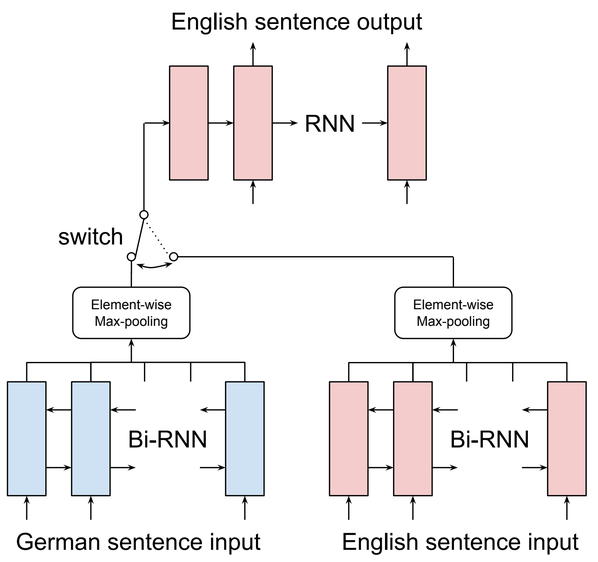
- Trained with:
  - `de-fr` bilingual data
  - `en-fr` bilingual data
- Non-English bilingual data is rare
  - except European languages

Autoencoding



- Trained with:
  - `de-en` bilingual data
  - `en` monolingual data
- Easy to obtain such data

Model & Training



- Each mini-batch has examples of
  - NMT: `de-en`
  - Autoencoding: `en-en` (identical input/output)
- Max-pooling: compress representations into a single vector
  - Better than: first/last state, average-pooling, attention
- Pre-trained cross-lingual word embedding: MUSE

Similarity of Sentence Embeddings

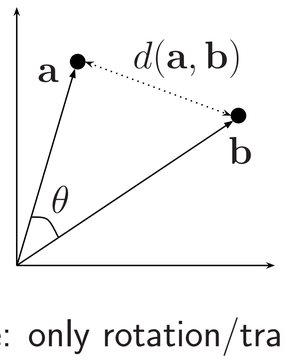
Problem: How to find similar sentences across languages?

Solution: Nearest neighbor search in the cross-lingual embedding space

Predefined functions

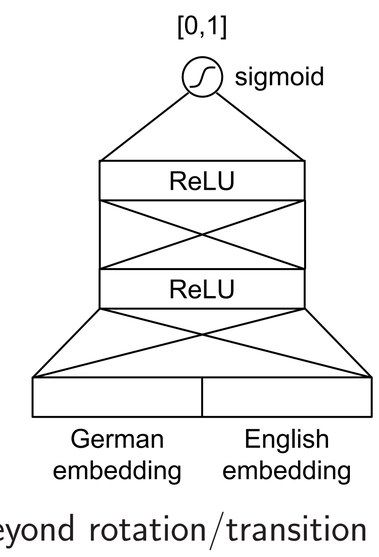
$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|$$



- Simple: only rotation/transition
- Assumption: sentence embeddings perfectly fit to vector geometry
  - Not guaranteed

Multilayer perceptron (MLP)

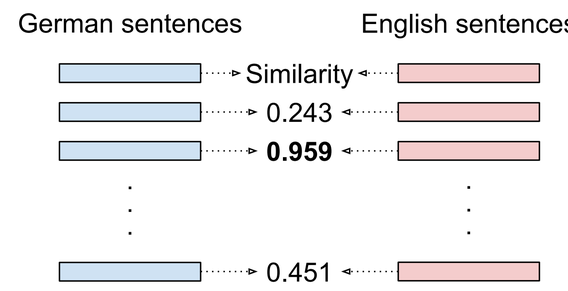


- Beyond rotation/transition
- Trained with tiny bilingual data
  - Optimized to a desired domain
- Compensates weak embeddings

Application: Parallel Corpus Filtering

Given: Noisy bilingual corpus (ParaCrawl, 36M sentence pairs)

Goal: Select only the sentence pairs with high similarity scores



WMT 2018 corpus filtering task

1. Select 10M-word amount
2. Train NMT model (3-layer Transformer) on selected data
3. Check BLEU scores on test sets

Scoring method		de-en BLEU [%] newstest2017	de-en BLEU [%] newstest2018
Random sampling		19.1	23.1
Pivot-based sentence embedding		26.1	32.4
NMT (de-en/en-de) + LM (de/en)		29.1	35.2
Bilingual sentence embedding (NMT + Autoencoding)	+ cos	23.0	28.4
	+ MLP	**29.2**	**35.4**

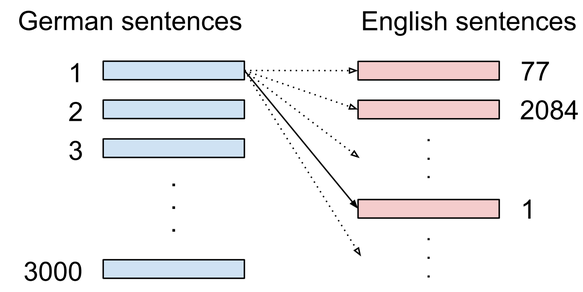
- 4-model combination of NMT + LM: long training/inference time
- Bilingual sentence embedding: much faster
  - Pivot-based: not practical (needs bilingual data with pivot language)
  - NMT + Autoencoding: best performance (MLP)

Application: Sentence Alignment Recovery

Given: Bilingual corpus with target side shuffled

- Sentence alignments are corrupted

Goal: Find the most similar target sentence for each source sentence



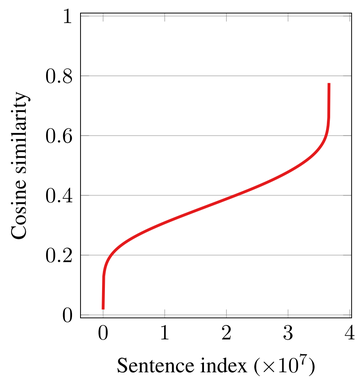
- Intrinsic evaluation for parallel corpus mining
- Error = $\frac{\#(\text{wrong alignments})}{\#(\text{total sentences})}$

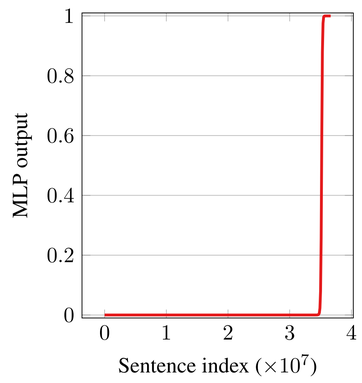
Similarity score		Error [%] newstest2018	Error [%] TED.tst2015	Time
Character-level edit distance (inverse)		37.4	54.6	5m
NMT model scores (de-en + en-de)		**1.7**	**13.3**	12h
Bilingual sentence embedding (NMT + Autoencoding)	+ cos	4.3	13.8	**27**s
	+ $-d$	53.8	61.6	2m
	+ MLP	89.9	72.6	1.5h

- NMT model scoring: too slow
- Bilingual sentence embedding: much faster, still decent performance (cos)
  - MLP: not suitable for this task (explanation below)

Cosine Similarity vs. Multilayer Perceptron

Question: Why MLP works well in filtering but not in alignment recovery?





- Sensitive to small errors
- Good for finding exact match
- High scores to all "okay" cases
- Good for keeping all useful pairs

Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS"), the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1, project "CoreTec"), and eBay Inc. The GPU cluster used for the experiments was partially funded by DFG Grant INST 222/1168-1. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

Paper Link