

Unsupervised Bilingual Lexicon Induction and Bitext Mining

Abdellah EL MEKKI

Cross-lingual Retrieval for Iterative Self-Supervised Training

Chau Tran
Facebook AI
chau@fb.com

Yuqing Tang
Facebook AI
yuqtang@fb.com

Xian Li
Facebook AI
xianl@fb.com


Jiatao Gu
Facebook AI
jgu@fb.com

Bilingual Lexicon Induction via Unsupervised Bilingual Construction and Word Alignment

Haoyue Shi *
TTI-Chicago
freda@ttic.edu

Luke Zettlemoyer
University of Washington
Facebook AI Research
lsz@fb.com

Sida I. Wang
Facebook AI Research
sida@fb.com



Motivations

- Parallel data is a key factor for building machine translation systems.
- Unsupervised mining of parallel will lead to unsupervised machine translation.
- Unsupervised Bitext mining + Strong word alignment system \Rightarrow high quality lexicons.



Cross-lingual Retrieval for Iterative Self-Supervised Training (CRISS)

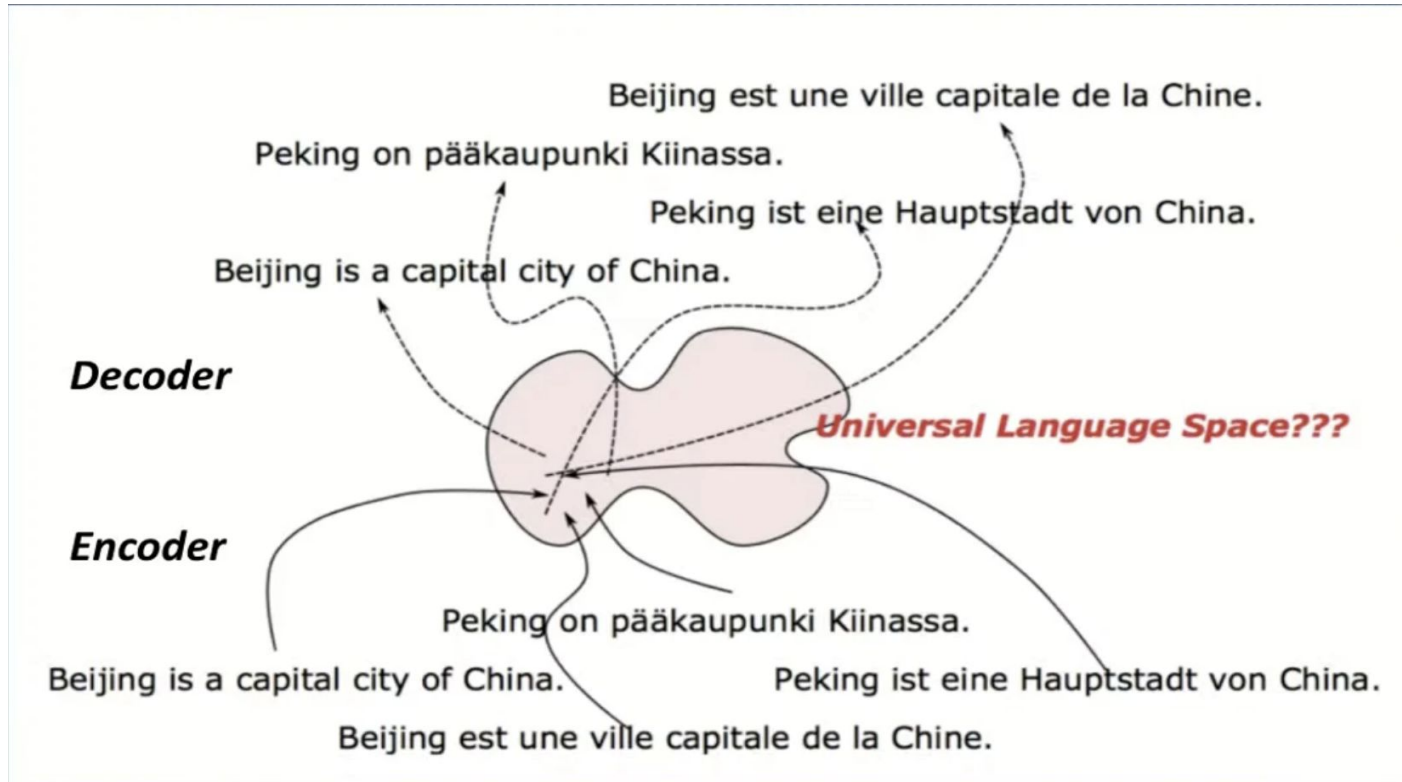


Summary

- CRISS is an iterative procedure that:
 - Mines for parallel sentences across languages using mBART.
 - Train a new better mBART using these mined sentences pairs.
 - Mines again for better parallel sentences, and repeats.



Multilingual Lingual Language Spaces Lead to Universal Language Space



mBART cross-lingual alignment

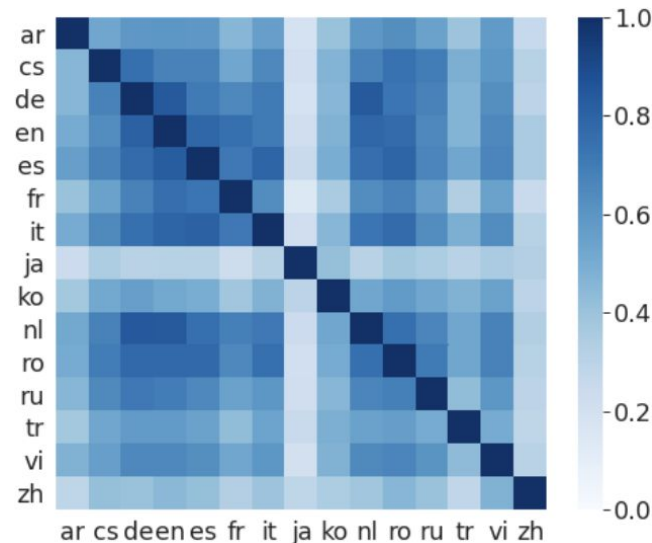
The pre-trained encoder tends to output similar representations across different languages without parallel supervision.

Task: Sentence retrieval

Data: TED58

For each language pair:

- Encode sentences with the pretrained mBART encoder.
- Use the average pooled last layer hidden states to search the nearest neighbor in the target language.
- Report top-1 accuracy.



57% (on average) vs 0.04% (random)

mBART cross-lingual alignment

The alignment gets stronger after fine-tuning on any pair

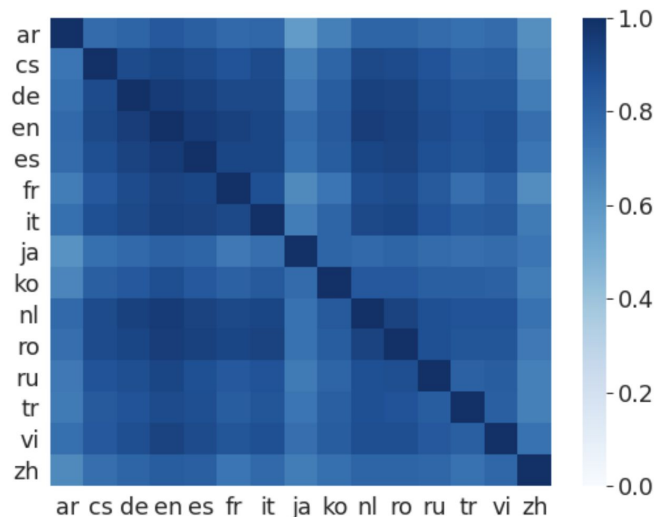
mBART is now fine-tuned on the bitext data of JA-En of TED58.

Task: Sentence retrieval

Data: TED58

After fine-tuning:

- The retrieval accuracy of all pairs gets improved.
- **When fine-tuning on bitext of any language, the model automatically learns to translate all languages because of the aligned representations.**



84% (on average) vs 57% (before)

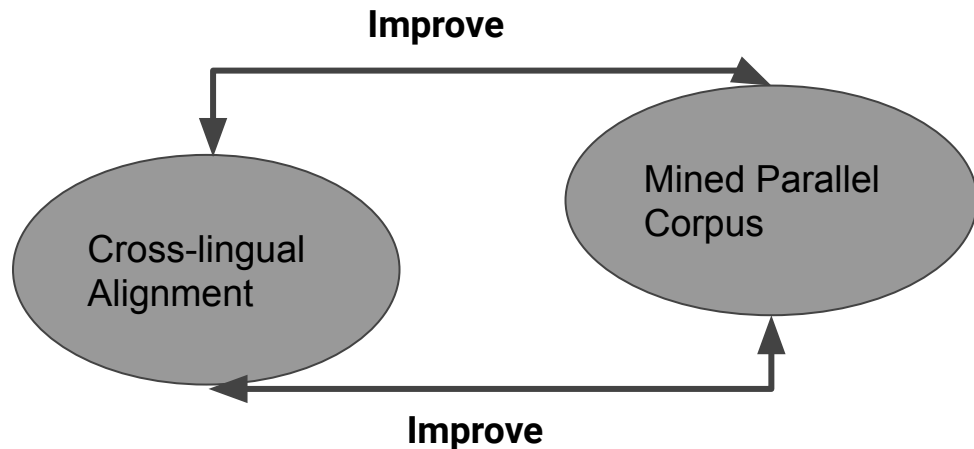
Q: How to reach better sentence retrieval with no parallel data?

A: Pseudo-Parallel data

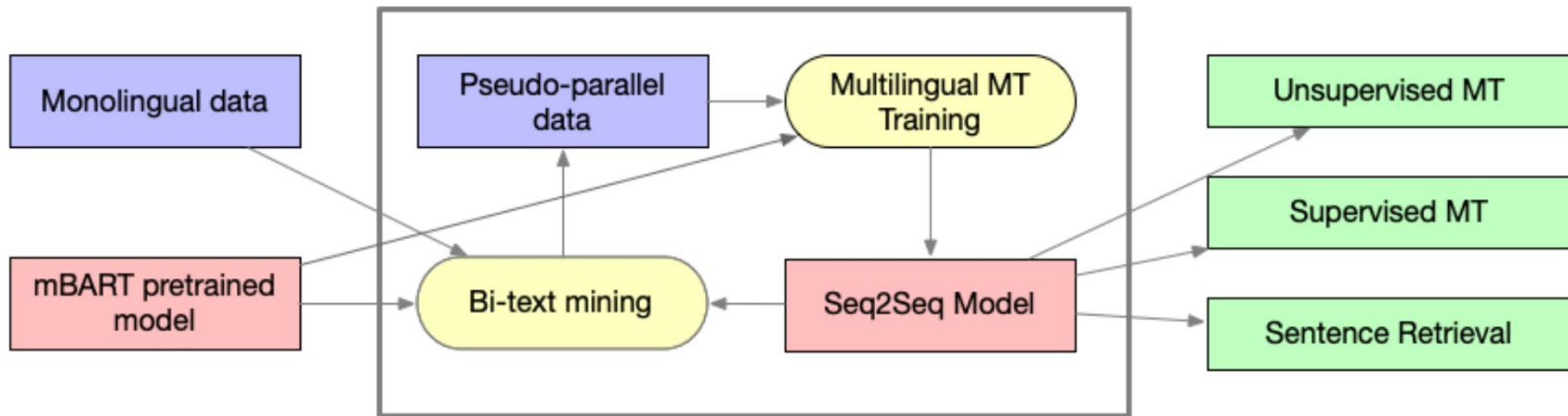


mBART cross-lingual alignment

- Real parallel data is replaced with pseudo parallel data mined by the model itself based on sentence retrieval.



Proposed framework



Unsupervised parallel data mining

- x and y are the vector representation of two sentences in two languages.
- Sentences are then encoded simply by extracting L2-normalized average-pooled encoder outputs.
-

$$\text{score}(x, y) = \frac{\cos(x, y)}{\sum_{z \in N_x} \frac{\cos(x, z)}{2k} + \sum_{z \in N_y} \frac{\cos(z, y)}{2k}}$$



Unsupervised parallel data mining

Algorithm 1 Unsupervised Parallel Data Mining

```
1: function MINE( $\Theta, D_i, D_j$ )
2:   Input: (1) monolingual data sets  $D_i$  and  $D_j$  for language  $i$  and  $j$  respectively, (2) a pretrained model  $\Theta$ ,
3:   Set  $k, M, \tau$  to be the desired KNN size, the desired mining size, and the desired minimum score threshold respectively
4:   for each  $x$  in  $D_i$ , each  $y$  in  $D_j$  do
5:      $x, y \leftarrow \text{Embed}(\Theta, x), \text{Embed}(\Theta, y)$ 
6:      $N_x, N_y \leftarrow \text{KNN}(x, D_j, k), \text{KNN}(y, D_i, k)$  ▷ Using FAISS [24]
7:   end for
8:   return  $D' = \{(x, y)\}$  where  $(x, y)$  are the top  $M$  pairs s.t.  $\text{score}(x, y) \geq \tau$  following Equation 2
9: end function
```

Iteratively mining and multilingual training

Algorithm 2 CRISS training

- 1: **Input:** (1) monolingual data from N languages $\{D_n\}_{n=1}^N$, (2) a pretrained mBART model Ψ , (3) total number of iterations T
 - 2: Initialize $\Theta \leftarrow \Psi, t = 0$
 - 3: **while** $t < T$ **do**
 - 4: **for** every language pairs (i, j) where $i \neq j$ **do**
 - 5: $D'_{i,j} \leftarrow \text{Mine}(\Theta, D_i, D_j)$ ▷ Algorithm 1
 - 6: **end for**
 - 7: $\Theta \leftarrow \text{MultilingualTrain}(\Psi, \{D'_{i,j} \mid i \neq j\})$ ▷ Note: Train from the initial mBART model Ψ
 - 8: **end while**
-

Results

Direction	en-de	de-en	en-fr	fr-en	en-ne	ne-en	en-ro	ro-en	en-si	si-en
CMLM [47]	27.9	35.5	34.9	34.8	-	-	34.7	33.6	-	-
XLM [13]	27.0	34.3	33.4	33.0	0.1	0.5	33.3	31.8	0.1	0.1
MASS [51]	28.3	35.2	37.5	34.9	-	-	35.2	33.1	-	-
D2GPO [33]	28.4	35.6	37.9	34.9	-	-	36.3	33.4	-	-
mBART [34]	29.8	34	-	-	4.4	10.0	35.0	30.5	3.9	8.2
CRISS Iter 1	21.6	28.0	27.0	29.0	2.6	6.7	24.9	27.9	1.9	6
CRISS Iter 2	30.8	36.6	37.3	36.2	4.2	12.0	34.1	36.5	5.2	12.9
CRISS Iter 3	32.1	37.1	38.3	36.3	5.5	14.5	35.1	37.6	6.0	14.5

Table 1: Unsupervised machine translation. CRISS outperforms other unsupervised methods in 9 out of 10 directions. Results on mBART supervised finetuning listed for reference.

Results

Language	ar	de	es	et	fi	fr	hi	it
XLMR [14]	47.5	88.8	75.7	52.2	71.6	73.7	72.2	68.3
mBART [34]	39	86.8	70.4	52.7	63.5	70.4	44	68.6
CRISS Iter 1	72	97.5	92.9	85.6	88.9	89.1	86.8	88.7
CRISS Iter 2	76.4	98.4	95.4	90	92.2	91.8	91.3	91.9
CRISS Iter 3	78.0	98.0	96.3	89.7	92.6	92.7	92.2	92.5
LASER (supervised) [6]	92.2	99	97.9	96.6	96.3	95.7	95.2	95.2

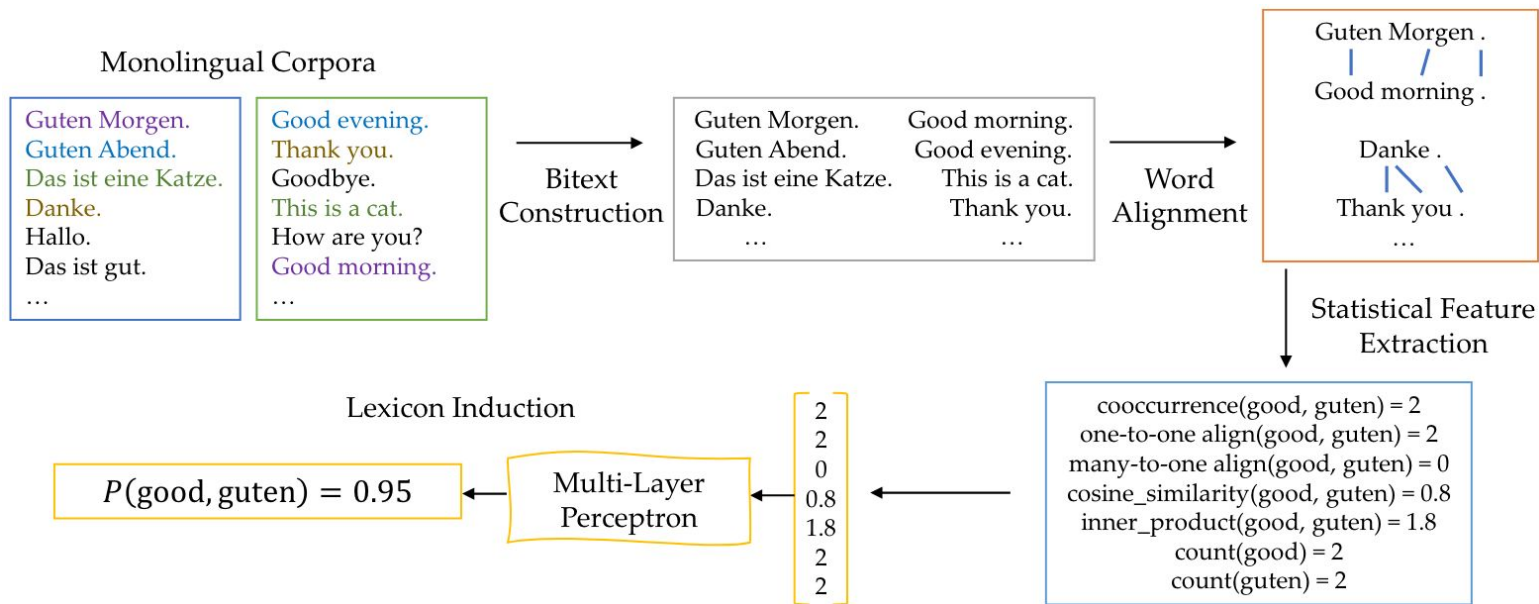
Language	ja	kk	ko	nl	ru	tr	vi	zh
XLMR [14]	60.6	48.5	61.4	80.8	74.1	65.7	74.7	68.3 (71.6)
mBART [34]	24.9	35.1	42.1	80	68.4	51.2	63.9	14.8
CRISS Iter 1	76.8	67.7	77.4	91.5	89.9	86.9	89.9	69
CRISS Iter 2	84.8	74.6	81.6	92.8	90.9	92	92.5	81
CRISS Iter 3	84.6	77.9	81.0	93.4	90.3	92.9	92.8	85.6
LASER (supervised) [6]	94.6	17.39	88.5	95.7	94.1	97.4	97	95

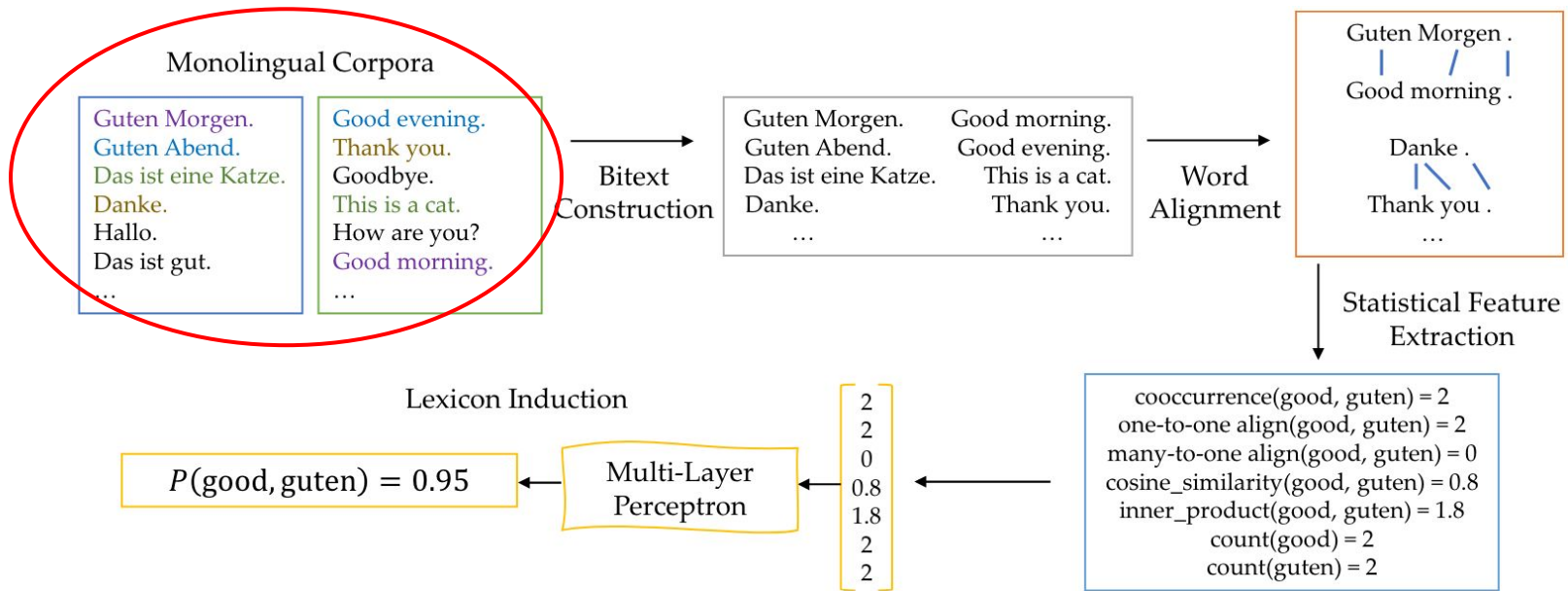
Table 2: Sentence retrieval accuracy on Tatoeba; XLMR results are the previous SOTA (except zh where mBERT is the SOTA). LASER is a supervised method listed for reference

Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment

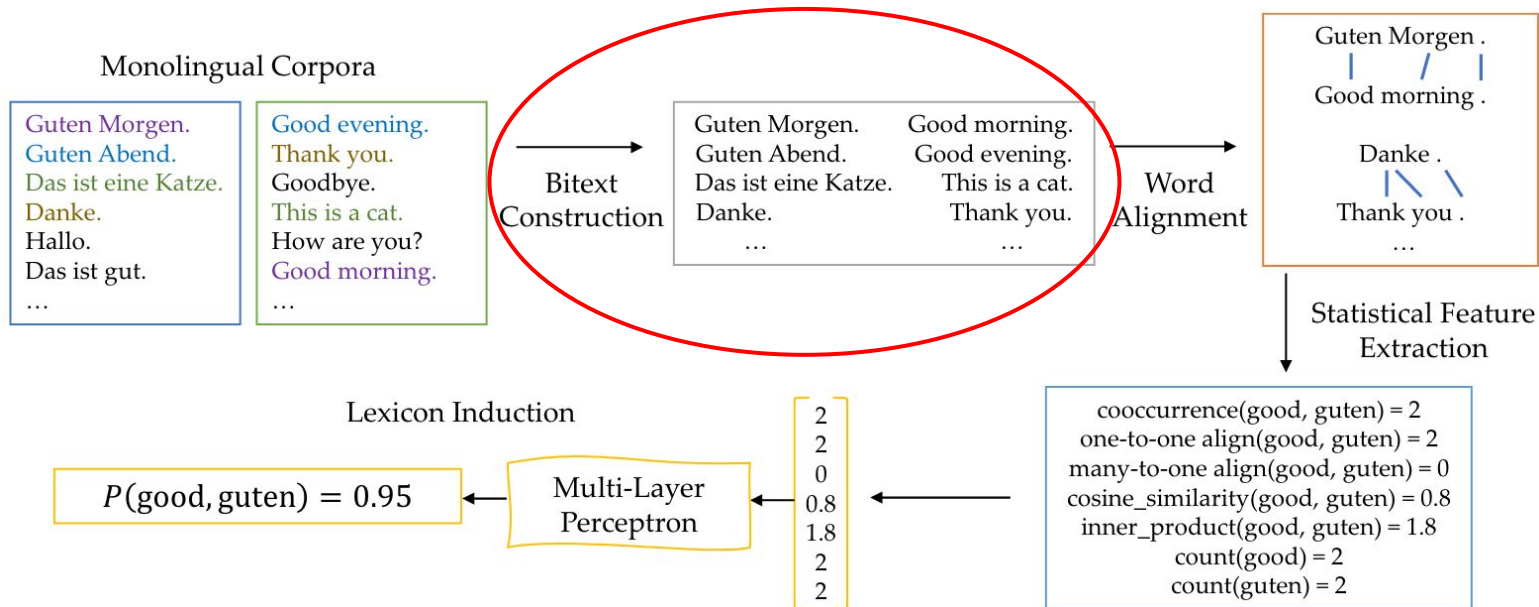


Proposed framework



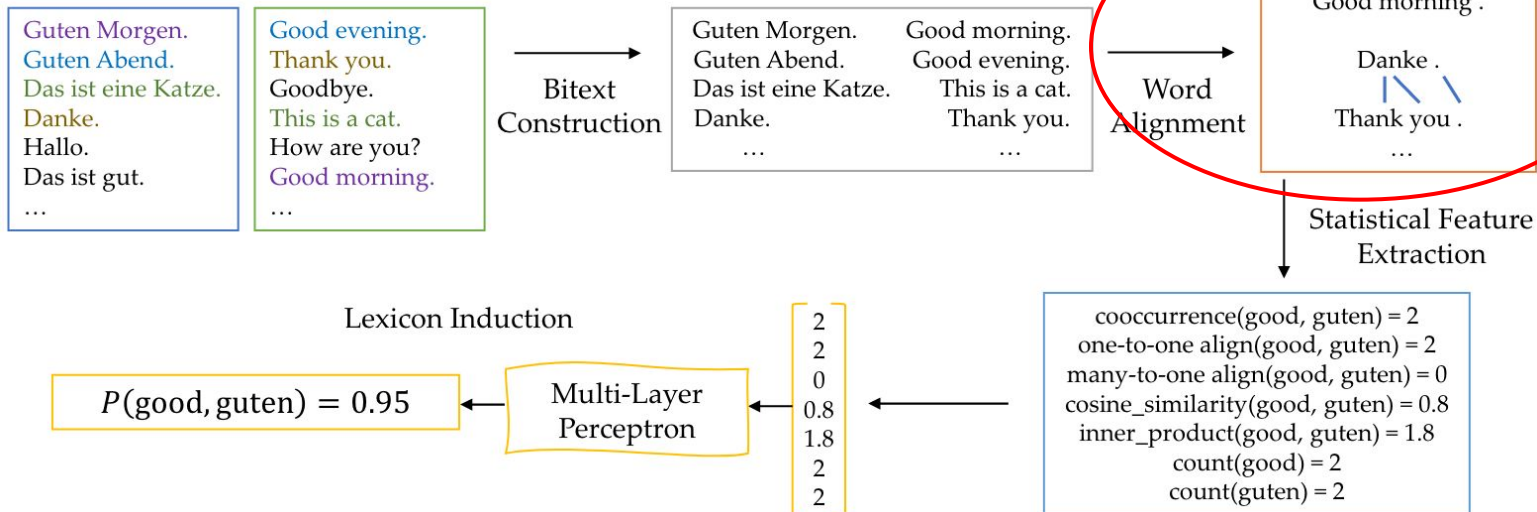


Monolingual Corpora: Separate monolingual corpora for source and target languages.



Bitext Construction: Using CRISS.

Monolingual Corpora



Word Alignment: Using SimAlign

SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings



SimAlign Method

- Alignments using multilingual language models such as mBERT and XLM-R.
- Cosine similarities matrix between every source token vector and every target token vector.

Der Pinguin Nils Olav wurde vom norwegischen König zum Ritter geschlagen

Pingvin Nils Olav Norvegiya qiroli tomonidan ritsar edi

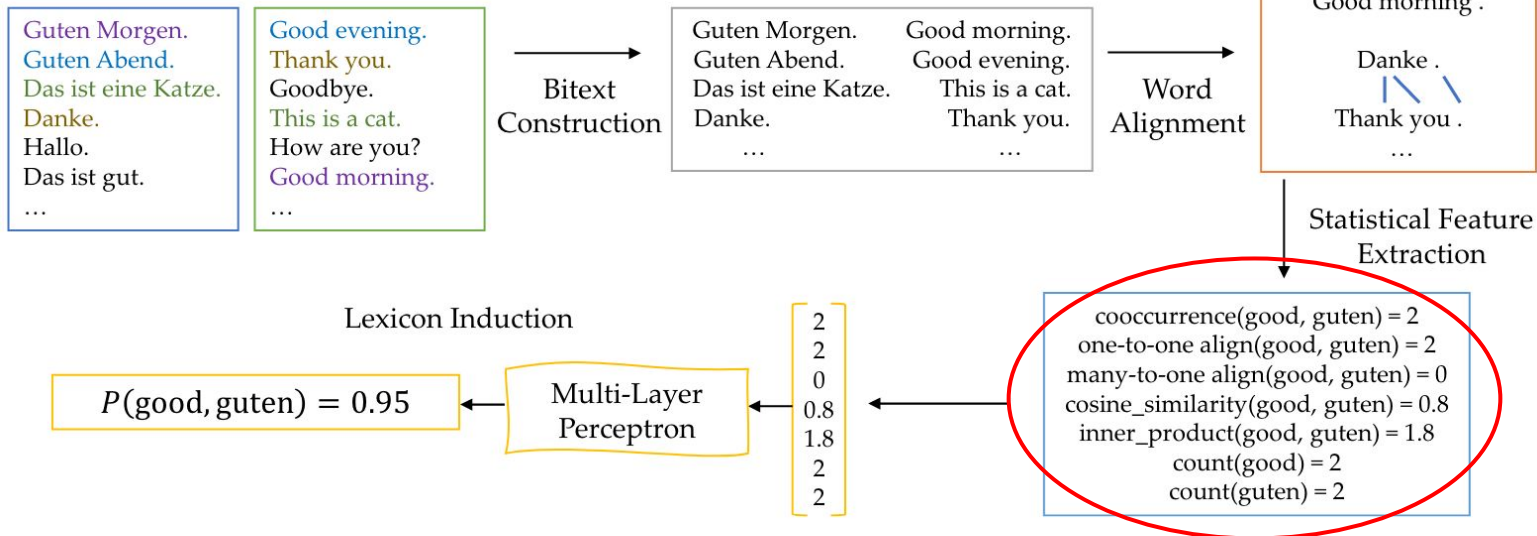


Sir Nils Olav III. です ペンギン knighted by el rey noruego

Nils Olav der Dritte is a penguin nominato cavaliere par un roi norvégien



Monolingual Corpora



Statistical feature extraction

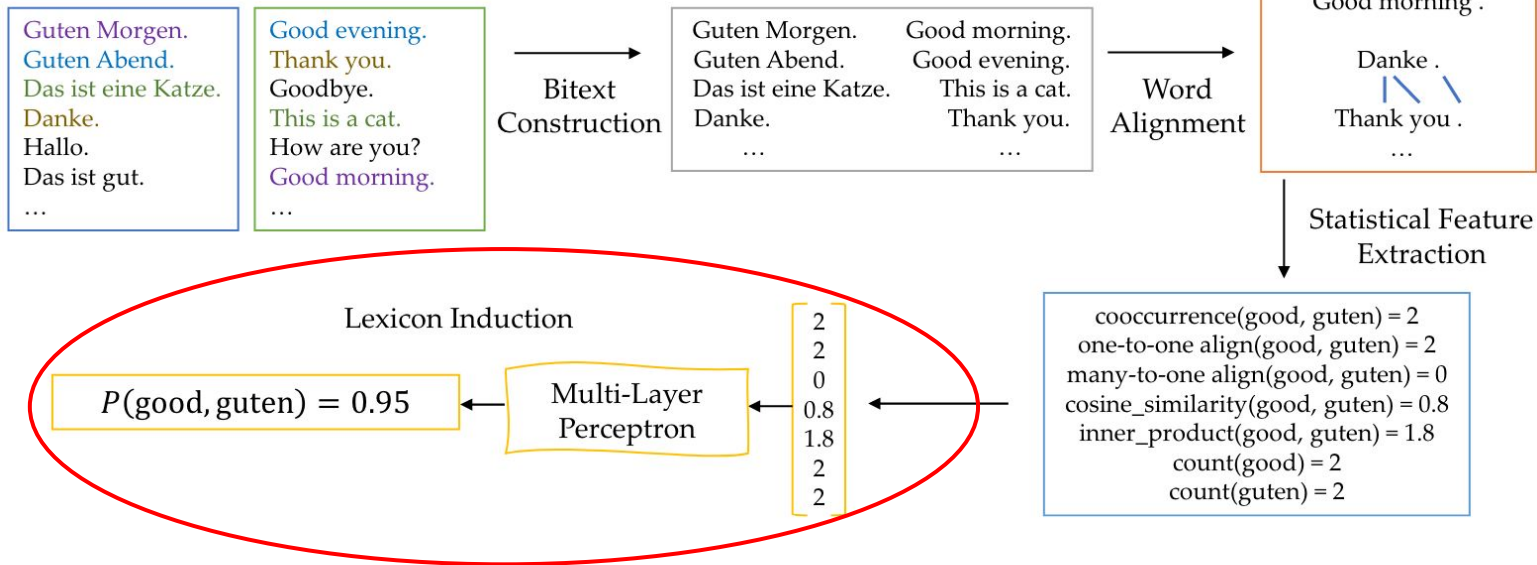
Statistical feature extraction

For a pair of words (s, t):

- Count of alignment:
- Count of co-occurrence
- The count of s in the source language and t in the target language.
- Non-contextualized word similarity: we feed the word type itself into CRISS, use the average pooling of the output subword embeddings, and consider both cosine similarity and dot-product similarity as features.



Monolingual Corpora



Lexicon induction using the MLP model

Results

Language Pair	Weakly-Supervised							Unsupervised		
	BUCC	VECMAP	WM	GEN	GEN-N	RTV	GEN-RTV	VECMAP	GEN	RTV
de-en	61.5	37.1	71.6	70.2	67.7	73.0	74.2	22.1	62.6	66.8
de-fr	76.8	43.2	79.8	79.1	79.2	78.9	83.2	27.1	79.4	80.3
en-de	54.5	33.2	62.1	62.7	59.3	64.4	66.0	33.7	51.0	56.2
en-es	62.6	45.3	71.8	73.7	69.6	77.0	75.3	44.1	60.2	65.6
en-fr	65.1	45.4	74.4	73.1	69.9	73.4	76.3	44.8	61.9	66.3
en-ru	41.4	29.2	54.4	43.5	37.9	53.1	53.1	24.6	28.4	45.4
en-zh	49.5	31.0	67.7	64.3	56.8	69.9	68.3	12.8	51.5	51.7
es-en	71.1	55.5	82.3	80.3	75.8	82.8	82.6	52.4	71.4	76.4
fr-de	71.0	46.2	82.1	80.0	78.7	80.9	81.7	46.0	76.4	77.3
fr-en	53.7	51.5	80.3	79.7	76.1	80.0	83.2	50.4	72.7	75.9
ru-en	57.1	44.8	72.7	61.1	59.2	72.7	72.9	42.1	51.8	68.0
zh-en	36.9	36.1	64.1	52.6	50.6	62.5	62.5	34.4	34.3	48.1
average	58.4	41.5	72.0	68.4	65.1	72.4	73.3	36.2	58.5	64.8

Table 1: F_1 scores ($\times 100$) on the BUCC 2020 test set (Rapp et al., 2020). The best number in each row is **bolded**.