Finding translations in unordered text using multilingual sentence representations

Filip Ginter[1,2] and Jenna Kanerva[1]

[1] TurkuNLP Group, University of Turku, Finland
[2] Silo.ai

turkunlp. github.io

Objective

- Gather parallel corpora without assuming any document-level information
- Work from two large, in principle unordered sets of sentences
- Hundreds of millions to units of billions range in terms of sentence count
- Reach for the parallel data which is not explicitly linked on document level

Talk context

- A 2016 manuscript that didn’t make it :^)
  - Quite much superseded by later work
- Talk expanded to include related work and make a broader overview of what’s out there
- Published results mostly on English-German, English-French
- We also include results on Finnish-English
  - Highly dissimilar language pair

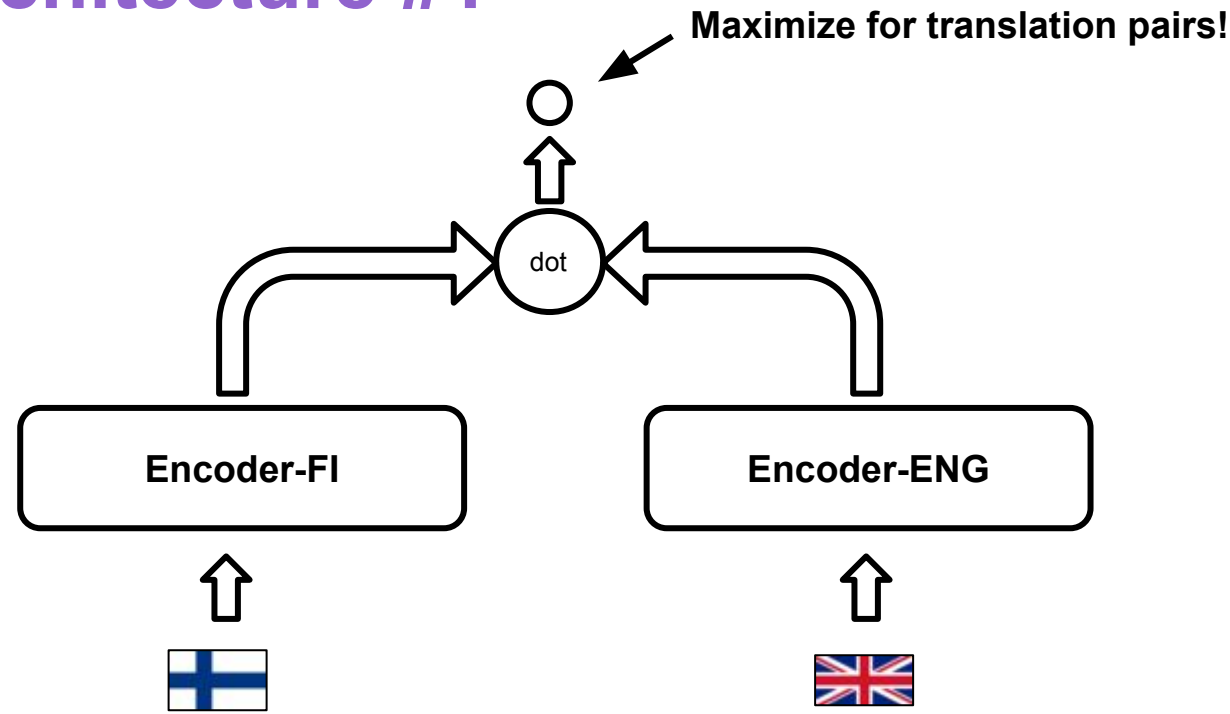
General approach

- Obtain cross-lingual sentence embeddings of all sentences in the monolingual corpora
  - Embedding - a vector representing a sentence
  - Sentences in a translation pair receive similar embeddings (in a vector comparison sense)
- Do all-against-all comparison of sentence pairs
- Sort by vector similarity
- For every sentence pick the most similar candidate in the other language
  - Similarity cut-off

Encoder

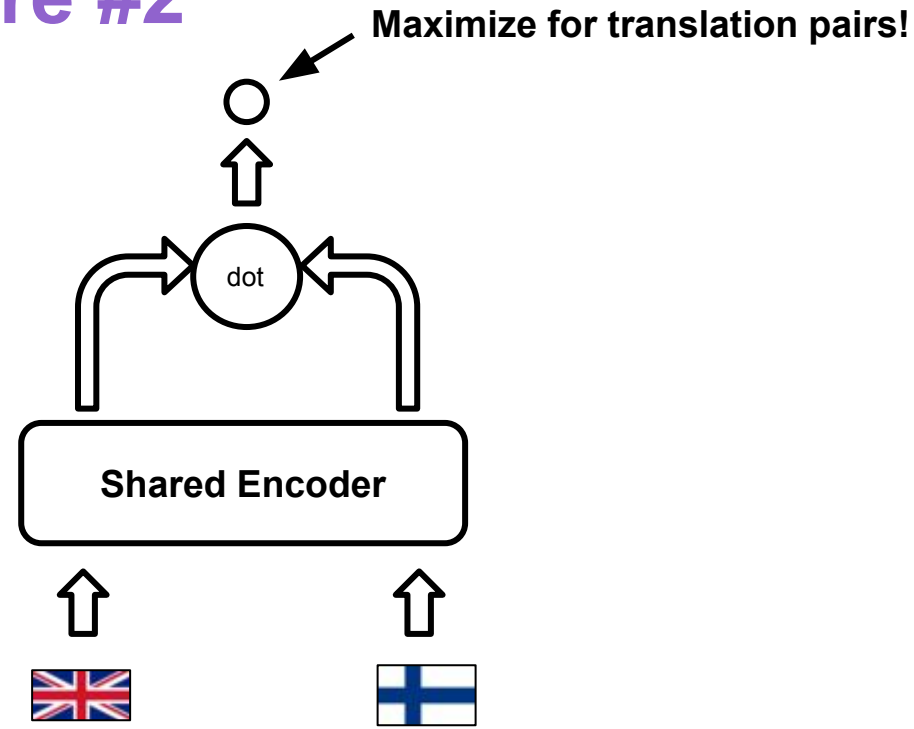
- Input: sequence of word vectors
  - Or sub-words such as BPE
- Output: **a single vector**
- Architecture - pick your favorite
- Seen in literature:
  - Deep averaging network
  - CNN + max pooling
  - (Bi)LSTM (final state or max pooling)

Architecture #1



UNIVERSITY OF TURKU

Architecture #2



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

Sampling negatives

- The negatives should not be too easy!
- Hard negatives based on initial sentence similarities (Guo et al. 2018)
  - Train a baseline model with random negative sampling
  - Use some of the high scoring candidates as hard negatives
  - Encoder forced to learn deep, not surface distinctions
  - Enough if done only for a part of the examples, rest with random negatives (saves processing time)

Sampling negatives - Evaluation

Negative Selection Approach	en-fr			en-es		
	P@1	P@3	P@10	P@1	P@3	P@10
Random Negative	34.83	47.99	61.20	44.89	58.13	70.36
Random Negative (Augmented)	36.51	49.07	61.37	47.08	59.55	71.34
(20) Hard Negative	48.90	62.26	73.03	54.94	67.78	78.06

Table 3: Precision at N (P@N) of target sentence retrieval on the UN corpus. Models attempt to select the true translation target for a source sentence from the entire corpus (11.3 million aligned sentence pairs.)

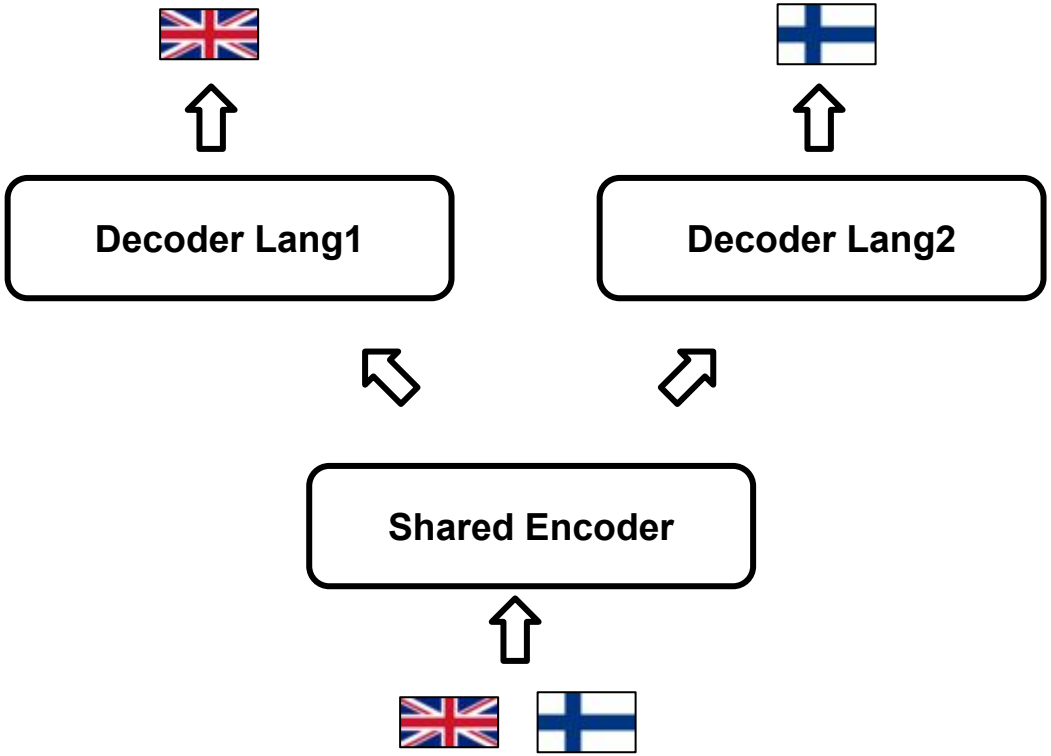
Guo et al. 2018. *Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. In Proceedings of the Third Conference on Machine Translation (WMT'18).* http://www.statmt.org/wmt18/pdf/WMT017.pdf

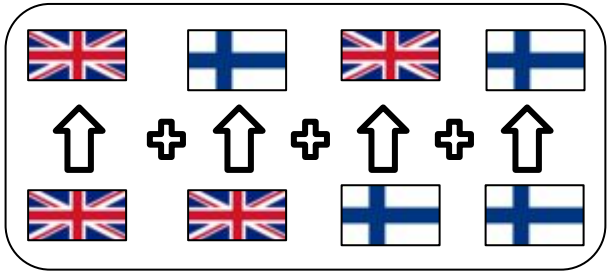
UNIVERSITY OF TURKU

Decoder

- Input: a single vector representing a sentence
- Output: the sentence itself
- Generated character / subword / word at a time
- Architecture seen in literature:
  - Left-to-right LSTM

Architecture #3



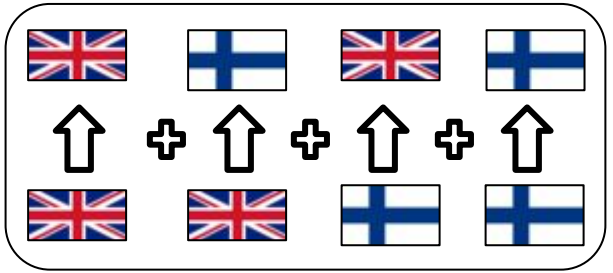


Training regime encourages cross-lingually comparable representations

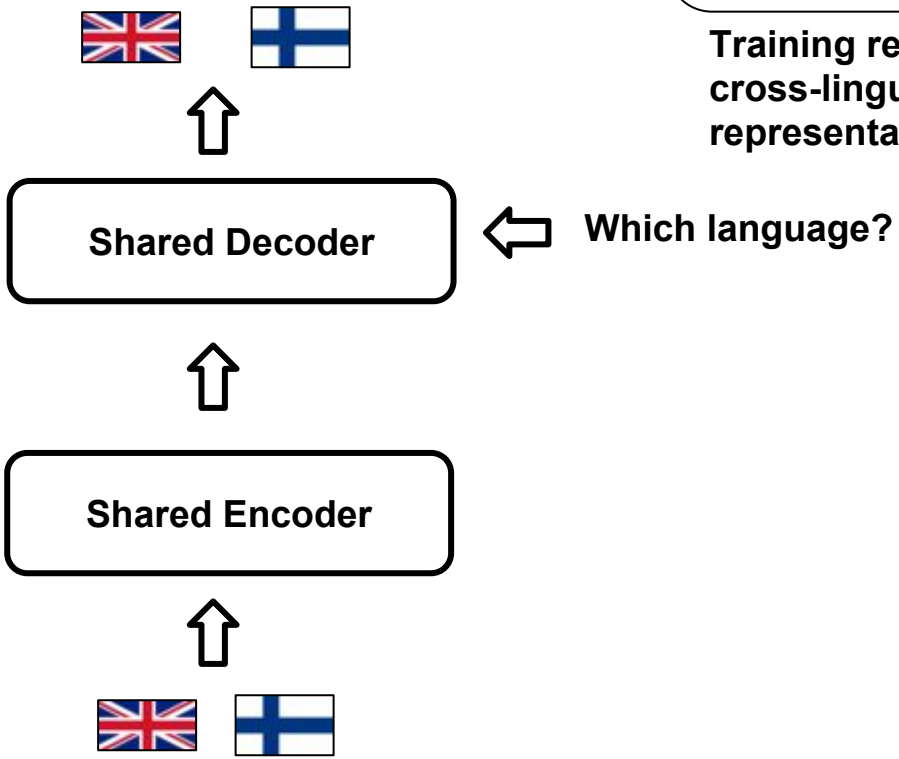
Constitutes a simple neural machine translation system

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

Parallel data construction

- Two monolingual corpora of substantial size
- Trained encoders encode all sentences
  - Two sets of sentence embeddings
  - Note: cannot embed sentence pairs - too heavy
- Compare the embeddings directly
  - All pairs in principle
  - If we had two corpora with 200M sentences each:
  - 200,000,000 x 200,000,000 = 💀

Similarity measure

- Dot product / Cosine
  - Scale of dot/cosine argued not to be consistent across different sentences
- Rescaled dot product (Guo et al. 2018)
  - Learn to rescale the dot value based on the source embedding
- Margin-based similarity (Artetxe and Schwenk 2018)
  - Instead of plain cosine, measure the margin between a given sentence pair and its closest candidates

Similarity measure - Evaluation

Margin funct.	Retrieval	EN-DE			EN-FR		
		P	R	F1	P	R	F1
Absolute (Cosine)	Forward	78.94	75.09	76.97	82.09	74.19	77.94
	Backward	78.96	73.07	75.90	77.24	72.24	74.66
	Intersection	84.89	80.76	82.78	83.60	78.33	80.88
	Max. score	83.14	77.18	80.05	80.86	77.53	79.16
Distance	Forward	94.79	94.09	94.44	91.05	**91.83**	91.44
	Backward	94.78	94.11	94.44	91.46	91.36	91.41
	Intersection	94.90	94.09	94.50	91.15	91.81	91.48
	Max. score	94.90	94.09	94.50	91.15	91.82	91.49
Ratio	Forward	95.18	94.39	94.79	92.37	91.29	91.83
	Backward	95.18	**94.42**	94.80	92.32	91.31	91.81
	Intersection	95.27	94.39	94.83	**92.43**	91.27	**91.85**
	Max. score	**95.28**	94.41	**94.84**	**92.43**	91.28	**91.85**

Table 2: Results on the BUCC mining task for different margin functions and retrieval strategies. We report the precision, recall and F1 score on the training set, used to optimize the filtering threshold for each variant.

Artetxe and Schwenk. *2018*.
Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *arXiv preprint arXiv:1811.01136.*
https://arxiv.org/abs/1811.01136

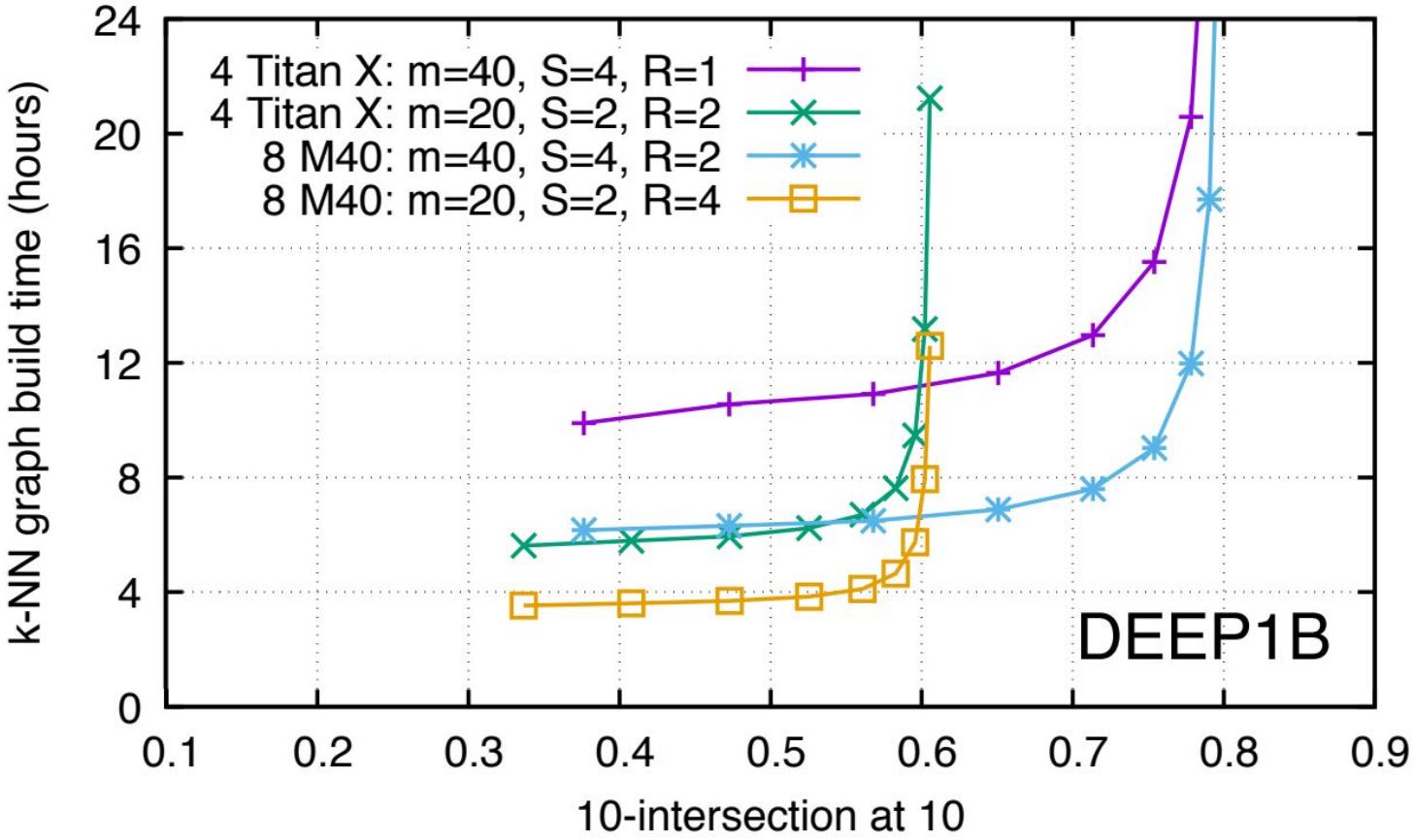
UNIVERSITY OF TURKU

Search at scale

- All pairs - costly
  - Our 2016 method on 170M vs 300M sentences
  - 50,000 CPU hours
  - Heavy filtering on sentence length
- Cluster and compare within nearest clusters
  - About 200h on our 170M vs 300M sentences
  - Still heavy filtered on sentence length

Search at scale

- FAISS https://github.com/facebookresearch/faiss
  https://arxiv.org/pdf/1702.08734.pdf
- Two-stage indexing
  a. Coarse k-means reduces to square-root
  b. Fine(r)-grained quantization thereafter within each coarse-grained cluster
- GPU implementation with good engineering
- Results in very fast search of nearest neighbors
- Efficient construction of k-NN graph
- Time/accuracy trade-off becomes a parameter



https://arxiv.org/pdf/1702.08734.pdf

UNIVERSITY OF TURKU

kNN graph

- Node: items
- Edge: link items to their k nearest neighbors

Figure 6: Path in the k-NN graph of 95 million images from YFCC100M. The first and the last image are given; the algorithm computes the smoothest path between them.

https://arxiv.org/pdf/1702.08734.pdf

Current SOTA: BUCC'18

- BUCC'18: Identifying parallel sentences in comparable corpora
  - Monolingual corpora with translation pairs inserted after the fact
  - Attempt to keep document structure "undamaged"
- Task: Given two sentence-split monolingual corpora, identify pairs of sentences that are translations of each other
- Pre/Rec/F1
- En-De/Fr/Ru/Zh

BUCC'18

	TRAIN				TEST			
	de-en	**fr-en**	**ru-en**	**zh-en**	**de-en**	**fr-en**	**ru-en**	**zh-en**
Azpeitia et al. (2017)	83.33	78.83	-	-	83.74	79.46	-	-
Grégoire and Langlais (2017)	-	20.67	-	-	-	20	-	-
Zhang and Zweigenbaum (2017)	-	-	-	43.48	-	-	-	45.13
Azpeitia et al. (2018)	84.27	80.63	80.89	76.45	85.52	81.47	81.30	77.45
Bouamor and Sajjad (2018)	-	75.2	-	-	-	76.0	-	-
Chongman Leong and Chao (2018)	-	-	-	58.54	-	-	-	56
Schwenk (2018)	76.1	74.9	73.3	71.6	76.9	75.8	73.8	71.6
Proposed method (Europarl)	**94.84**	**91.85**	-	-	**95.58**	**92.89**	-	-
Proposed method (UN)	-	90.75	**90.92**	**91.04**	-	-	**92.03**	**92.57**

Table 3: F1 scores on the BUCC mining task. Our proposed method uses the *ratio* margin function with *maximum score* retrieval, and the filtering threshold was optimized on the training set.

Artetxe & Schwenk (2018)

UNIVERSITY OF TURKU

Current SOTA: UN Corpus

- Reconstructing the United Nations parallel corpus
- Shuffled parallel corpus
- Task: Given a sentence, find its translation pair
- 11M sentences per language
- Each sentence has a pair

	EN-FR	EN-ES
Guo et al. (2018)	48.90	54.94
Proposed method	**83.27**	**85.78**

Table 4: Results on UN corpus reconstruction (P@1)

Artetxe & Schwenk (2018)

UNIVERSITY OF TURKU

Parallel data filtering

- Most common application of these techniques
- ParaCrawl filtering especially
- ~2 points BLEU score improvement on English-German

	DATA	BLEU	
		tok	detok
Wu et al. (2016)	wmt	26.3	-
Gehring et al. (2017)	wmt	26.4	-
Vaswani et al. (2017)	wmt	28.4	-
Ahmed et al. (2017)	wmt	28.9	-
Shaw et al. (2018)	wmt	29.2	-
Ott et al. (2018)	wmt	29.3	28.6
Ott et al. (2018)	wmt+pc	29.8	29.3
Edunov et al. (2018)	wmt+nc	35.0	33.8
Proposed method	pc	31.2	30.5
	wmt+pc	31.8	31.1

Table 6: Results on English-German newstest2014 in comparison to previous work. *wmt* for WMT parallel data (excluding ParaCrawl), *pc* for ParaCrawl, and *nc* for monolingual News Crawl with back-translation.

Artetxe and Schwenk. 2018

Comparable corpora construction

- Wikipedia article pairs
- English-French
- Does not scale up

Training Data	Model	BLEU	Sentences
Europarl		21.5	500,000
+Full	BiRNN	26.2 (+4.7)	1,987,769
	Baseline	25.4 (+3.9)	1,292,514
+Top500k	BiRNN	25.0 (+3.5)	1,000,000
	Baseline	24.9 (+3.4)	1,000,000

Table 3: BLEU scores obtained on the newstest2013 test set. Sentences is the number of sentences used to train the SMT systems. The Europarl row is the baseline SMT system trained on 500k sentences pairs from the Europarl corpus.

Gregoire and Langlais (2017)

Web-scale runs

Threshold	#Sents	BLEU		
		Mined alone	Eparl + mined	All + mined
baseline	-	-	21.87	25.06
0.25	1.0M	4.18	**22.32**	**25.07**
0.26	1.5M	5.17	22.09	-
0.27	1.9M	5.92	21.97	-
0.28	2.5M	6.48	22.29	25.03
0.29	3.3M	6.01	22.10	-
0.30	4.3M	7.77	22.24	-

Table 6: BLEU scores when training on the mined data only, adding it (at different thresholds) to the human translated training corpus (Eparl+NC) and to our best system using filtered Common Crawl.

CommonCrawl news.
English-German. Schwenk (2018)

Training data	BLEU score
Europarl (baseline)	13.45
Europarl + 200K	14.08
Europarl + 400K	14.09
Europarl + 600K	14.21
Europarl + 1M	14.22
Europarl + 2M	14.35
Europarl + 3M	14.19

Web crawl. English-Finnish.
Kanerva et al. (2016, unpublished)

UNIVERSITY OF TURKU

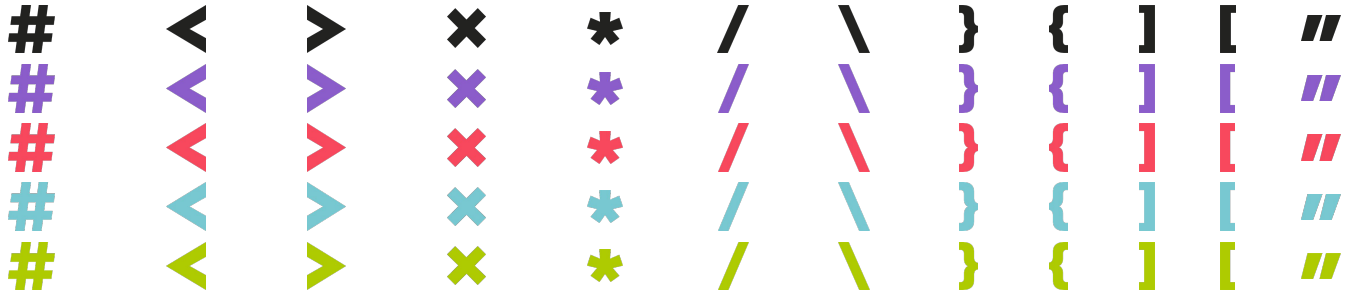
Summary

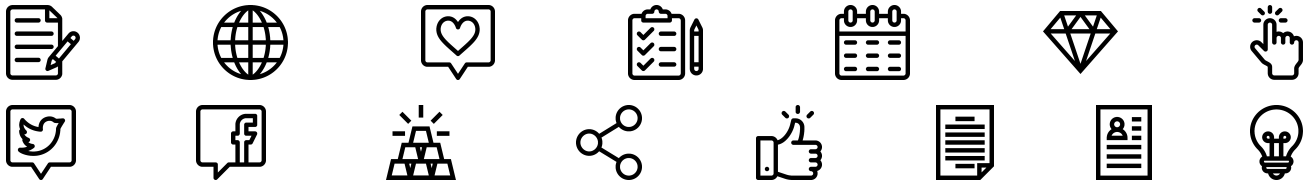
- Several methods for parallel data extraction without assuming document-level information
- Mostly applied to further filtering noisy parallel or comparable corpora
- Only a handful of studies on large monolingual corpora not enriched in translation pairs
  - Results quite weak
  - Very difficult to draw broader conclusions (yet)

References

- **Guo et al. 2018**, Effective Parallel Corpus Mining using Bilingual Sentence Embeddings https://arxiv.org/abs/1807.11906 *(introduces hard negatives + rescaling)*
- **Grégoire and Langlais 2017**, A Deep Neural Network Approach To Parallel Sentence Extraction https://arxiv.org/abs/1709.09783 *(pairwise classifier, also used in BUCC task)*
- **Espana-Bonet et al. 2017**, An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification https://arxiv.org/pdf/1704.05415.pdf
- **Schwenk and Douze 2017**, Learning Joint Multilingual Sentence Representations with Neural Machine Translation http://aclweb.org/anthology/W17-2619 *(introduces encoder-decoder approach)*
- **Schwenk 2018**, Filtering and Mining Parallel Data in a Joint Multilingual Space https://arxiv.org/abs/1805.09822
- **Artetxe and Schwenk 2018**, Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings https://arxiv.org/abs/1811.01136 *(introduces margin-based scoring, s.o.t.a.)*

Graafiset elementit, ikonit ja logot





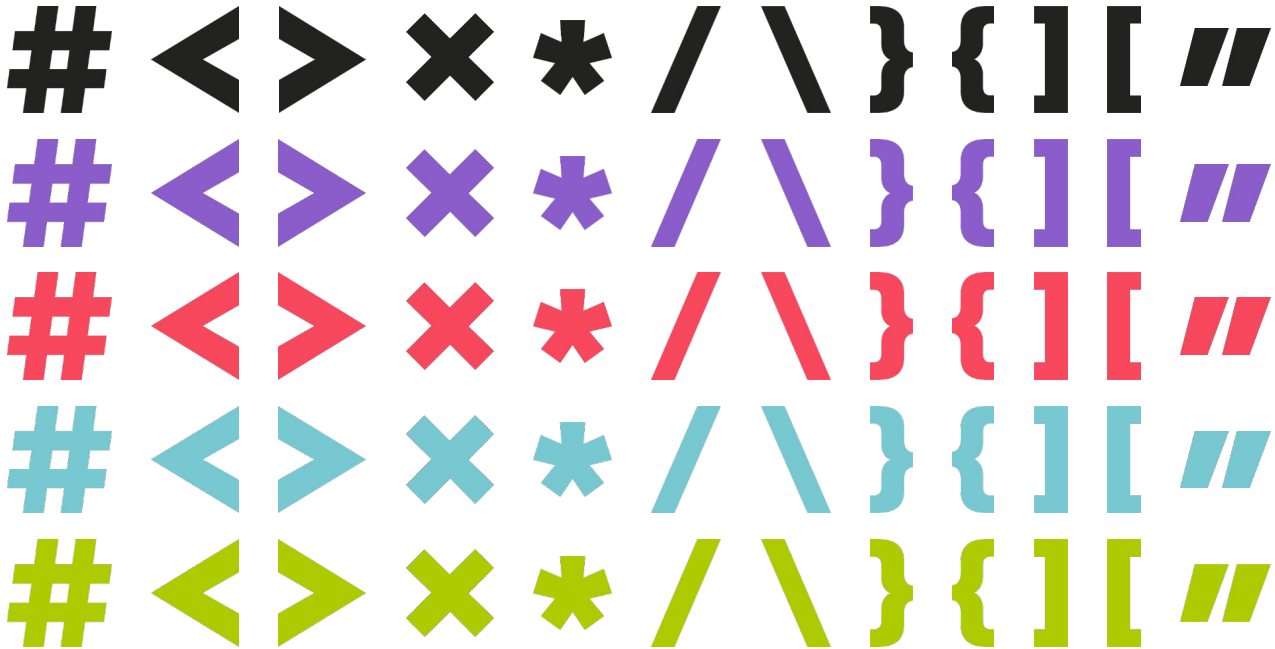
UNIVERSITY OF TURKU

TURUN YLIOPISTO

UNIVERSITY OF TURKU

TURUN YLIOPISTO

Graphic elements



UNIVERSITY OF TURKU