Exercise 07 – Neural Machine Translation

Chair XII for Natural Language Processing

Prof. Dr. Goran Glavaš
Fabian David Schmidt
Benedikt Ebing

Bitext Mining

What is bitext mining?

- Automatically retrieving parallel sentences (i.e., sentences that are translations of each other) from large corpora (e.g., the web).

Why is bitext mining important for NMT?

- Parallel sentences are *crucial* for the training of multilingual sentence encoders and neural machine translation.
- For higher resource languages, comparably large parallel corpora exist (e.g., United Nations Parallel Corpus or Europarl).
- Specifically for lower resource languages, there are less efforts to manually construct parallel corpora, creating the need for automatically mined parallel data.

How does CCMatrix work?

1. Text extraction from the CCNet corpus
   - Parse from JSON
   - Deduplicate sentences
   - Perform language identification (to filter noise)
2. Get all sentence embeddings using a sentence encoder (storing them in a compressed from using FAISS library)
3. For all pairs of sentences for two languages
   - Compute the margin-based criterion for both directions
   - Build the union of the forward and backward direction, sort the candidates and omit source or target sentences that are already used
   - Apply a threshold (hyperparameter) to decide whether two sentences are mutual translations.

NMT with Large Language Models

1. LLMs vs. supervised NMT models

Language Family	Direction	Translation Performance (BLEU / COMET)									
		XGLM-7.5B	OPT-175B	Falcon-7B	LLaMA2-7B	LLaMA2-7B-Chat	ChatGPT	GPT-4	M2M-12B	NLLB-1.3B	Google
Indo-Euro-Germanic (8)	X⇒Eng	18.54 / 70.09	34.65 / 83.71	27.37 / 67.40	37.28 / 84.73	34.82 / 84.25	45.83 / 89.05	48.51 / **89.48**	42.72 / 87.74	46.54 / 88.18	**51.16** / 89.36
	Eng⇒X	9.16 / 50.21	18.89 / 71.97	13.19 / 52.93	22.78 / 76.05	19.44 / 73.63	36.34 / 87.83	40.64 / 88.50	37.30 / 86.47	38.47 / 87.31	**45.27** / **89.05**
Indo-Euro-Romance (8)	X⇒Eng	31.11 / 79.67	38.93 / 87.75	34.06 / 84.40	41.10 / 88.10	37.84 / 87.80	45.68 / 89.61	**47.29** / **89.74**	42.33 / 88.31	46.33 / 88.99	35.69 / 89.66
	Eng⇒X	21.95 / 69.08	24.30 / 79.07	20.02 / 70.36	27.81 / 82.05	25.50 / 79.67	41.35 / **89.00**	**44.47** / 88.94	42.98 / 87.56	43.48 / 88.12	37.10 / 88.77
Indo-Euro-Slavic (12)	X⇒Eng	13.20 / 64.24	20.83 / 74.80	13.15 / 57.34	34.00 / 84.90	30.94 / 83.90	39.27 / 87.74	41.19 / 88.15	35.87 / 85.97	39.23 / 87.08	**43.61** / **88.18**
	Eng⇒X	6.40 / 43.28	8.18 / 54.45	4.34 / 35.73	20.24 / 76.30	16.14 / 69.75	32.61 / 87.90	36.06 / 89.15	35.01 / 86.43	36.56 / 88.74	**42.75** / **90.05**
Indo-Euro-Indo-Aryan (10)	X⇒Eng	8.68 / 63.93	1.20 / 49.37	1.40 / 45.22	6.68 / 62.63	4.29 / 60.29	25.32 / 84.14	37.30 / 87.79	17.53 / 69.66	40.75 / 88.80	**45.66** / **89.43**
	Eng⇒X	4.76 / 40.99	0.14 / 31.85	0.13 / 25.84	1.61 / 35.92	1.24 / 34.74	16.50 / 68.43	21.35 / 73.75	14.44 / 65.32	34.04 / 82.55	**39.04** / **82.78**
Indo-Euro-Other (11)	X⇒Eng	7.32 / 55.29	7.80 / 59.60	7.04 / 51.59	14.27 / 69.87	11.46 / 67.64	29.54 / 84.52	37.29 / 86.76	22.38 / 77.47	36.16 / 86.81	**41.68** / **88.29**
	Eng⇒X	4.51 / 40.60	3.10 / 40.04	3.38 / 34.64	5.00 / 44.09	4.83 / 43.73	22.81 / 77.33	28.45 / 80.94	19.71 / 74.90	31.65 / 85.82	**38.54** / **87.44**
Austronesian (6)	X⇒Eng	16.19 / 78.80	25.60 / 78.03	18.62 / 75.36	26.70 / 80.21	24.39 / 80.39	39.95 / 87.29	46.81 / 88.65	31.84 / 84.76	45.41 / 87.85	**50.68** / **88.89**
	Eng⇒X	10.01 / 73.14	10.68 / 64.97	8.56 / 60.89	14.59 / 74.80	13.29 / 74.88	30.17 / 86.36	34.66 / 87.68	27.03 / 86.83	37.17 / 88.82	**40.74** / **89.34**
Atlantic-Congo (14)	X⇒Eng	6.67 / 62.00	9.17 / 57.59	6.98 / 0.56	8.76 / 57.72	9.01 / 57.86	19.86 / 79.63	28.27 / 83.42	10.55 / 76.43	**32.20** / 84.00	23.55 / **85.44**
	Eng⇒X	2.52 / 54.93	1.60 / 34.15	1.89 / 0.34	2.45 / 34.17	3.09 / 38.13	8.91 / 75.26	13.70 / 77.79	6.53 / 75.79	**21.99** / 79.95	16.77 / **80.89**
Afro-Asiatic (6)	X⇒Eng	6.70 / 54.51	5.93 / 52.90	4.87 / 38.62	10.41 / 57.72	8.65 / 58.27	20.84 / 70.39	30.48 / 78.76	10.00 / 66.98	32.69 / 82.99	**36.14** / **84.47**
	Eng⇒X	2.07 / 41.48	1.40 / 41.86	1.40 / 27.64	3.22 / 43.04	3.07 / 43.39	13.57 / 67.60	19.36 / 75.56	7.83 / 68.86	26.08 / 82.84	**31.00** / **83.78**
Turkic (5)	X⇒Eng	7.43 / 61.69	7.89 / 62.47	4.15 / 33.11	9.51 / 65.95	8.88 / 66.15	24.64 / 84.04	31.73 / 86.90	10.25 / 58.52	32.92 / 87.51	**37.78** / **88.53**
	Eng⇒X	3.48 / 40.32	2.58 / 44.80	1.75 / 20.00	3.28 / 39.65	3.09 / 41.97	17.13 / 74.77	20.96 / 78.50	10.87 / 68.21	30.17 / 88.47	**36.54** / **89.38**
Dravidian (4)	X⇒Eng	8.04 / 61.95	0.89 / 44.01	1.18 / 24.29	2.65 / 53.17	1.52 / 52.95	20.26 / 82.00	33.10 / 86.91	10.26 / 63.77	39.07 / 88.42	**43.17** / **89.10**
	Eng⇒X	5.30 / 48.15	0.02 / 32.51	0.03 / 15.31	0.56 / 34.03	0.58 / 35.65	12.34 / 64.74	18.60 / 75.15	6.85 / 62.25	37.33 / 86.32	**44.16** / **87.75**
Sino-Tibetan (3)	X⇒Eng	9.35 / 58.60	9.32 / 65.32	16.59 / 72.34	18.35 / 74.45	16.88 / 74.20	21.36 / 78.52	27.74 / 84.48	11.09 / 71.35	30.88 / 86.50	**35.68** / **87.66**
	Eng⇒X	10.14 / 74.16	2.57 / 54.73	10.74 / 66.74	12.24 / 65.99	9.06 / 65.07	19.92 / 76.04	22.81 / 81.11	10.42 / 73.82	16.85 / 80.74	**32.40** / **88.52**
Other (14)	X⇒Eng	9.71 / 60.43	10.10 / 60.78	5.37 / 47.38	16.00 / 71.15	14.25 / 70.35	25.59 / 82.48	32.62 / 86.21	25.53 / 81.53	35.06 / 86.86	**36.95** / **87.93**
	Eng⇒X	8.42 / 51.57	3.82 / 46.85	1.73 / 29.73	8.19 / 53.20	7.14 / 52.12	20.26 / 74.31	24.04 / 79.59	23.29 / 77.80	28.54 / 85.84	**34.34** / **87.82**

2. Typical errors of LLMs (removed from the paper)

Type I: Off-target translation	
Source	Prime Minister Stephen Harper has agreed to send the government's 'Clean Air Act' to an all-party committee for review, before its second reading, after Tuesday's 25 minute meeting with NDP leader Jack Layton at the PMO.
Hypothesis	Baş nazir Stephen Harper, NDP lideri Jack Layton ile PMO'daki 25 dakikalık görüşmesinin ardından hükümetin 'Temiz Hava Yasası'nı ikinci okumadan önce tüm parti komitesine gözden geçirmesi için göndermeyi kabul etti.
Reference	Baş Nazir Stiven Harper Layihe İdareetme Ofisinde (LIO) Yeni Demokratik Partiya (YDP) lideri Cek Layton ile çerşenbe axşamı 25 deqiqelik iclasdan sonra ikinci oxunuşdan evvel hökumetin Temiz Hava Aktınıñezerden keçirilmesi üçün bütün partiyaların üzv olduğu komiteye göndermeyi qebul etdi
Error	ChatGPT fails to translate English into turkish, but into Azerbaijani.
Type II: Hallucination	
Source	Huwobe illimu kimiya be fasaha jodi hefti do kujjeji ko jomirawo tagi ha asamaji do mari alaka be luttukal ko taginga ha asamaji fu bana no mari alaka be kala ko be tagi andigel pat.
Hypothesis	The study of chemistry and language is important in our modern world, and we must work hard to improve our knowledge in both areas. Let us all strive to learn and improve ourselves in these fields so that we can make a positive impact on society.
Reference	The scientists were able to conclude that the dark matter affect other dark matter in the same way regular matter does.
Error	ChatGPT generates highly pathological translations that are completely unrelated with the source sentence.
Type III: Monotonic translation	
Source	They found the Sun operated on the same basic principles as other stars: The activity of all stars in the system was found to be driven by their luminosity, their rotation, and nothing else.
Hypothesis	他们发现太阳按照同样的基本原理运作，就像其他星体一样：系统中所有星体的活动均由它们的亮度、旋转驱动，且仅有这些因素
Reference	他们发现，太阳和其他恒星的运行原理是一样的：星系中所有恒星的活跃度都完全决定于它们的光度和自转
Error	ChatGPT translates English sentence word-by-word, lacking effective word-reordering.

3. Data Leakage

- Large language (and instruction-tuned) models might be trained on publicly available evaluation data
- Might inflate the performance on certain evaluation datasets
- Particularly an issue for commercial models that do not open-source their training corpora which hinders fair comparison

4. Importance of templates for prompting

In-context Template	Deu-Eng	Eng-Deu	Rus-Eng	Eng-Rus	Rus-Deu	Deu-Rus	Average
reasonable templates:							
<X>=<Y>	37.37	**26.49**	29.66	22.25	17.66	**17.31**	**25.12**
<X> \n Translate from [SRC] to [TGT]: \n <Y>	37.95	26.29	29.83	20.61	17.56	15.93	24.70
<X> \n Translate to [TGT]: \n <Y>	37.69	25.84	**29.96**	19.61	17.44	16.48	24.50
<X> \n [TGT]: <Y>	29.94	17.99	25.22	16.29	12.28	11.71	18.91
<X> is equivalent to <Y>	23.00	4.21	17.76	9.44	8.14	9.84	12.07
<X>\n can be translated to\n <Y>	37.55	**26.49**	29.82	22.14	17.48	16.40	24.98
[SRC]: <X> \n [TGT]: <Y>	16.95	8.90	14.48	6.88	7.86	4.01	9.85
unreasonable templates:							
<X>$<Y>	37.77	26.43	29.53	20.99	17.72	17.27	24.95
<X> \n Translate from [TGT] to [SRC]: \n <Y>	**38.18**	26.21	29.85	20.35	**17.75**	16.63	24.83
<X> \n Compile to [TGT]: \n <Y>	37.39	26.35	29.68	19.91	17.52	16.15	24.50
<X> \n [SRC]: <Y>	27.86	16.69	24.41	18.16	11.98	12.60	18.62
<X> is not equivalent to <Y>	23.50	3.92	16.90	7.80	8.06	9.23	11.57
<X> \n can be summarized as \n <Y>	37.46	26.24	29.42	**22.62**	17.68	17.15	25.10
[SRC]: <X> \n [SRC]: <Y>	19.03	8.21	15.96	6.37	7.57	4.40	10.26

5. Importance of Parallel Data

- Traditionally, encoder-decoder MT used large amount of parallel data
- Uprise of LLMs as strong MT questioned the use of parallel data
- ALMA: Continual pretraining of English-centric LLM on multiple monolingual corpora, followed by instruction tuning on a few-thousand of high-quality instances
- Tower: Continual pretraining on monolingual (2/3) and parallel data(1/3), followed by instruction tuning on translation-related tasks

➔LLMs for MT use smaller fraction of parallel data (10%-50%)

➔Parallel data still very helpful for low-resource languages (sample efficient)

➔Parallel data improves performance at larger scale

➔Parallel data often used at the end of the training pipeline (at the end of pretraining or in instruction-tuning)

Additional Exercise

Decoding

We are given the following probabilities for tokens {BOS, A, B, C, D} at timesteps $\{T\}_{i=0}^{3}$ for a text generation model:

	T = 0	T = 1	T = 2	T = 3
BOS	0.140	0.257	0.248	0.149
A	0.391	0.096	0.402	0.336
B	0.197	0.341	0.267	0.358
C	0.271	0.305	0.083	0.157

Compute the probabilities for the decoded sequence for both (i) greedy decoding and (ii) beam search with width k=3. Show your intermediate steps.

Greedy Decoding

- Pick the character with the highest probability: the decoded string is ABAB

$$P(ABAB) = 0.391 \times 0.341 \times 0.402 \times 0.358 = 0.0191$$

Beam Search

- Use logarithm to avoid underflow

	T = 0	T = 1	T = 2	T = 3
BOS	-1.966	-1.357	-1.393	-1.902
A	-0.939	-2.343	-0.912	-1.091
B	-1.625	-1.075	-1.322	-1.027
C	-1.304	-1.186	-2.460	-1.847

Beam Search – Step 0

BOS	-1.966
A	**-0.939**
B	**-1.625**
C	**-1.304**

Beam Search – Step 1

	BOS	A	B	C
A	-0.939 + -1.357 = **-2.296**	–3.283	**–2.015**	**–2.126**
B	-1.625 + -1.357 = -2.982	–3.969	–2.700	–2.812
C	-1.304 + -1.357 = -2.661	–3.647	–2.379	–2.490

Beam Search – Step 2

	BOS	A	B	C
A->B	–3.408	**–2.927**	–3.337	–4.504
A->C	–3.519	**–3.038**	–3.448	–4.615
A->BOS	–3.689	**–3.208**	–3.618	–4.756

Beam Search – Step 3

	BOS	A	B	C
A->B->A	-4.830	*-4.017*	**-3.953**	-4.778
A->C->A	-4.942	-4.128	*-4.065*	-4.889
A->BOS->A	-5.113	-4.300	-4.236	-5.061

Compute BLEU-4

- *Reference:* The quick brown fox jumps over the lazy dog
- *Hypothesis:* The fast brown fox leaps over the lazy dog

Unigrams

- *Reference:* [The, quick, brown, fox, jumps, over, the, lazy, dog]
- *Hypothesis:* [The, fast, brown, fox, leaps, over, the, lazy, dog]
- *Matching:* [The, brown, fox, over, the, lazy, dog]
- *Precision:* $p_1 = \frac{7}{9}$

Bigrams

- *Reference:* [(The, quick), (quick, brown), **(brown, fox)**, (fox, jumps), (jumps, over), **(over, the)**, **(the, lazy)**, **(lazy, dog)**]
- *Hypothesis:* [(The, fast), (fast, brown), **(brown, fox)**, (fox, leaps), (leaps, over), **(over, the)**, **(the, lazy)**, **(lazy, dog)**]
- *Matching:* [(brown, fox), (over, the), (the, lazy), (lazy, dog)]
- *Precision:* $p_2 = \frac{4}{8}$

Trigrams

- *Reference:* [(The, quick, brown), (quick, brown, fox), (brown, fox, jumps), (fox, jumps, over), (jumps, over, the), **(over, the, lazy)**, **(the, lazy, dog)**]
- *Hypothesis:* [(The, fast, brown), (fast, brown, fox), (brown, fox, leaps), (fox, leaps, over), (leaps, over, the), **(over, the, lazy)**, **(the, lazy, dog)**]
- *Matching:* [(over, the, lazy), (the, lazy, dog)]
- *Precision:* $p_3 = \frac{2}{7}$

4-grams

- *Reference:* [(The, quick, brown, fox), (quick, brown, fox, jumps), (brown, fox, jumps, over), (fox, jumps, over, the), (jumps, over, the, lazy), **(over, the, lazy, dog)**]
- *Hypothesis:* [(The, fast, brown, fox), (fast, brown, fox, leaps), (brown, fox, leaps, over), (fox, leaps, over, the), (leaps, over, the, lazy), **(over, the, lazy, dog)**]
- *Matching:* [(over, the, lazy, dog)]
- *Precision:* $p_4 = \frac{1}{6}$

Geometric Mean of Precisions

- $Geometric\ mean = \prod_{n=1}^{4}(p_n)^{\frac{1}{4}} = \exp(\frac{1}{4} \times \sum_{n=1}^{4} \log(p_n))$
- Plug numbers in:
  - $= \exp(\frac{1}{4} \times (\log\left(\frac{7}{9}\right) + \log\left(\frac{1}{2}\right) + \log\left(\frac{2}{7}\right) + \log(\frac{1}{6})))$
  - $= \exp(\frac{1}{4} \times log\left(\frac{7}{9} \times \frac{1}{2} \times \frac{2}{7} \times \frac{1}{6}\right))$
  - $= \exp\left(\frac{1}{4} \times \log\left(\frac{1}{54}\right)\right)$
  - ≈ 0.369

Brevity Penalty

- Reference length: $r = 9$
- Hypothesis length: $c = 9$
- $BP = \begin{cases} 1 & if\ c > r \\ e^{1-\frac{r}{c}} & if\ c \leq r \end{cases}$
- ➔ $BP = 1$

Final BLEU-4

- $BLEU4 = BP \times Geometric\ Mean$
- $\approx 1 \times 0.369 \approx 0.369$