

CUNI Submission for the MT4All Task

SIGUL 2022

Ivana Kvapilíková, Ondřej Bojar

Charles University
Institute of Formal and Applied Linguistics

24 - 25th June 2022
Marseille, France



Introduction

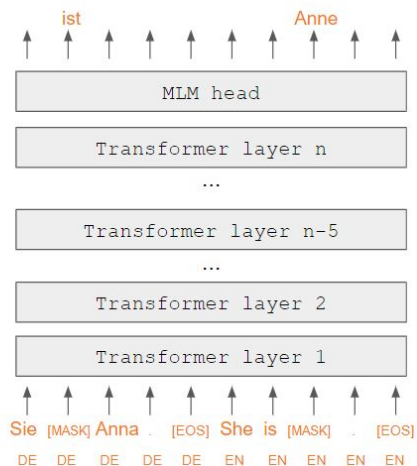
- Translation from English to Ukrainian, Kazakh and Georgian
- Legal domain
- Monolingual resources only
- In-domain data (120-270k sentences for kk, ka, en and 8.6M sentences for uk)
- Out-of-domain data (large Oscar dataset with 6-7M sentences for kk, ka and >20M sentences for uk, en)

Training Strategy

- **Preprocessing**
 - Language ID (fasttext), sentence splitting (nltk)
- **Cross-lingual pretraining** (Lample et al., 2019)
 - Multilingual masked language model jointly trained for en, ka, kk, uk
 - Shared subword dictionary of 50k items
- **Baseline fine-tuning**
 - Online back-translation
 - Denoising
- **Our fine-tuning**
 - Online back-translation
 - Denoising
 - Supervised MT training on *mined pseudo-parallel corpus*

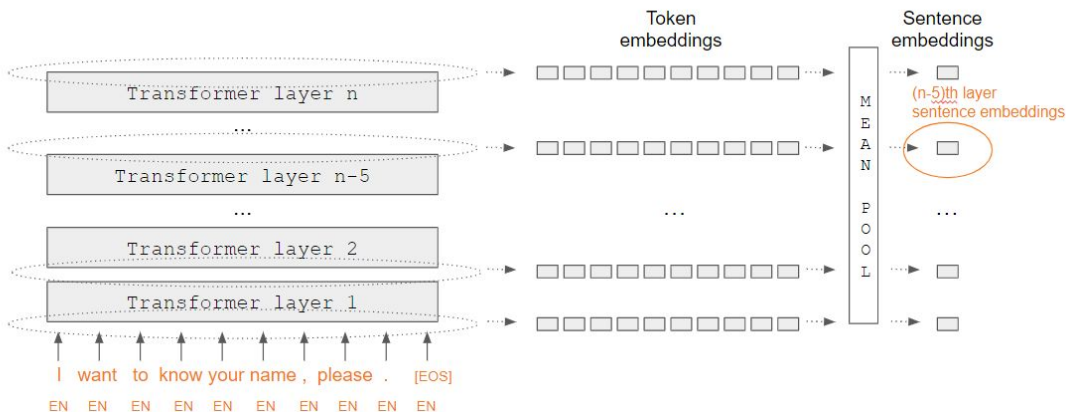
Parallel Data Mining (1/2)

- Unsupervised sentence embeddings (Kvapilikova et al., 2020)
- Representations from pretrained multilingual language model (XLM-100)
- Fine-tuned on 20k synthetic sentence pairs (English-German)



Parallel Data Mining (1/2)

- Unsupervised sentence embeddings (Kvavilikova et al., 2020)
- Representations from pretrained multilingual language model (XLM-100)
- Fine-tuned on 20k synthetic sentence pairs (English-German)



Parallel Data Mining (2/2)

- Nearest neighbour search in the embedding space to find pairs
- Margin-based cosine similarity for scoring candidates (Conneau et al, 2018)

	en-ka	en-kk	en-uk
monolingual	22.4M x 6.3M	22.4M x 7.6M	22.4M x 9.7M
mined (all)	8.8M	4.7M	21.0M
mined (selected)	400K	300K	600K
mined (cleaned)	230K	169K	496K

Ukrainian	English
Encyclopædia Britannica, англ.	Encyclopædia Britannica, Inc.
№ 538.	Number 538.
Це 100%.	This 100%.
Все життя.	Nice life.
Їй було 35.	He was 31.
Свій!	Sure!
І він відмінно працює!	It works perfectly!
Це надзвичайно цікава історія.	It's an extraordinarily beautiful work.
Ці компанії раніше вже були включені [...]	Search features have been added into [...]
Одним з таких є приватна медична практика.	One of those is analytic continuation.
6 місяців назад вона народила дитину.	And two years ago she had another healthy baby boy.
Сьогодні стає зрозуміло, що боротьба з COVID-19 триватиме не один рік.	Today it is clear that the fight against COVID-19 will last more than one year.
Людське тіло містить від 55% до 78% води.	Human beings are made up of 50 – 86% water.

Excerpt from the pseudo-parallel corpus for Ukrainian – English

System Description

- **Model architecture**
 - Same as Lample et al. (2019)
 - 6-layer Transformer
 - 1024 hidden dimension
 - GELU activations
 - Language embeddings
- **Training**
 - Dropout 0.2
 - Vocabulary size 30k
- **Inference**
 - Beam search with beam size 4

Results

Flores dev set

		en-ka	en-kk	en-uk
1	MT-BT-DN	1.9	1.7	6.7
2	XLM + BT-DN	1.6	1.4	4.5
3	XLM + MT-BT-DN	3.4	2.7	9.0
4	(3) + BT	4.1	3.7	10.6
5	(4) + legal BT	4.1	3.8	8.8

Official test set

		en-ka	en-kk	en-uk
1	MT-BT-DN	-	-	-
2	XLM + MT-BT-DN	-	-	-
3	(2) + BT	13.8	7.7	27
4	(3) + legal BT	13.8	9.4	28.1
	Baseline	12	6.4	20.8

Legend

XLM crosslingual masked LM pretraining

MT supervised MT training on mined corpus

BT online back-translation

DN denoising

Baseline hybrid SMT+NMT system (Artetxe et al., 2019)

Conclusion

- Significant MT improvement achieved by using additional **pseudo-parallel data** for training
- Mined samples of **various quality**
- Positive effect of **in-domain fine-tuning**

Thank you for your attention!

Any questions?

A large, dark blue, diagonal shape that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.

References

Mikel Artetxe, Holger Schwenk. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. ACL 2019.

Mikel Artetxe, Gorka Labaka, Eneko Agirre. An Effective Approach to Unsupervised Machine Translation. ACL 2019.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, Ondřej Bojar. Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining. ACL SRW 2020.

Guillaume Lample, Alexis Conneau. Cross-lingual Language Model Pretraining. NeurIPS 2019.