Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation

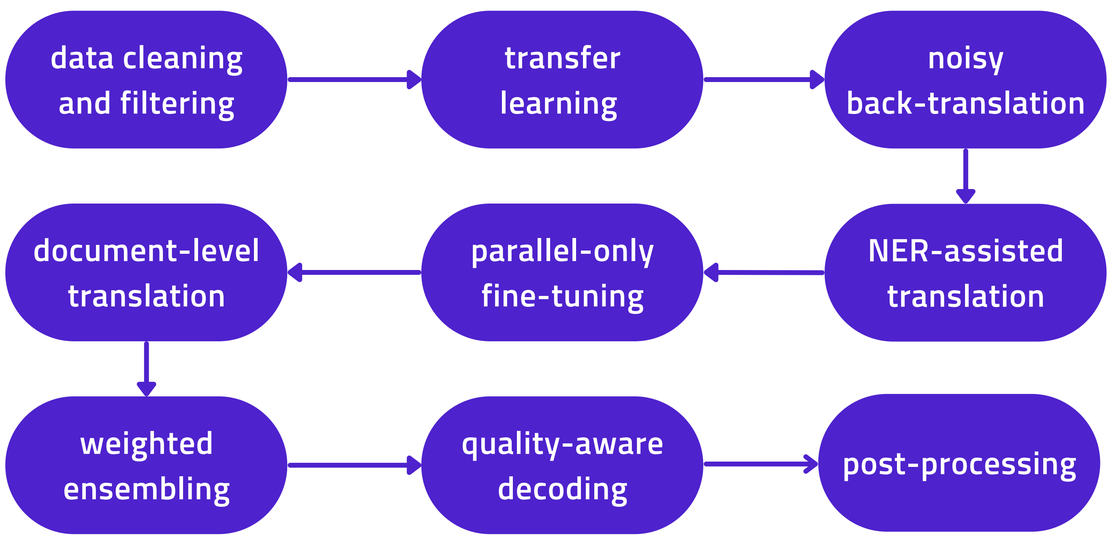
Artur Nowakowski[12], Gabriela Pałka[13], Kamil Guttmann[12], Mikołaj Pokrywka[12]

[1]Adam Mickiewicz University, Poznań, Poland
[2]Poleng, Poznań, Poland
[3]Applica.ai, Warsaw, Poland
{artur.nowakowski,gabriela.palka}@amu.edu.pl, {kamgut,mikpok1}@st.amu.edu.pl

The Story Behind

The poster presents Adam Mickiewicz University's (AMU) submissions to the constrained track of the WMT 2022 General MT Task. We participated in the Ukrainian ↔ Czech translation directions – a low-resource translation scenario between closely related languages. The models were trained using only the data provided by the shared task organizers. Our solution is composed of multiple translation quality augmentation techniques, including NER-assisted translation and quality-aware decoding. According to the evaluation results, our solutions rank first in both translation directions.

The Path Taken



NER-Assisted Translation

→ The named entity recognition was applied to assign appropriate source factors to words in the text, supporting the translation process.

Source factors mapping

factor	named entity	Czech (Slavic BERT model)	Ukrainian (Stanza NER module)
p0	---	---	---
p1	person (PER)	+	+
p2	location (LOC)	+	+
p3	organization (ORG)	+	+
p4	miscellaneous (MISC)	-	+
p5	product (PRO)	+	-
p6	event (EVT)	+	-

→ Models were trained in two settings: concatenation and sum. In the first setting, the factor embedding had a size of 16 and was concatenated with the token embedding. In the second setting, the factor embedding was equal to the size of the token embedding (1024) and was summed with it.

```
Hlavní|p0 inspektor|p0 organizace|p0 RSPCA|p3 pro|p0 Nový|p2 Jižní|p2 Wales|p2
David|p1 O'Shannessy|p1 televizi|p0 ABC|p5 sdělil|p0 ,|p0 že|p0 dohled|p0 nad|p0
jatky|p0 a|p0 jejich|p0 kontroly|p0 by|p0 měly|p0 být|p0 v|p0 Austrálii|p2
samozřejmostí|p0 .|p0

_Hlavní|p0 _inspektor|p0 _organizace|p0 _R|p3 SP|p3 CA|p3 _pro|p0 _Nový|p2 _Jižní|p2
_Wales|p2 _David|p1 _O|p1 '|p1 S|p1 han|p1 ness|p1 y|p1 _televizi|p0 _A|p5 BC|p5
_sdělil|p0 ,|p0 _že|p0 _dohled|p0 _nad|p0 _ja|p0 tky|p0 _a|p0 _jejich|p0 _kontroly|p0
_by|p0 _měly|p0 _být|p0 _v|p0 _Austrálii|p2 _samozřejmost|p0 í|p0 .|p0
```

An example of a sentence tagged with NER source factors before and after subword encoding.

Document-Level Translation

→ Our document-level translation approach is based on a simple sentence concatenation method.

→ We used parallel document-level corpora, as well as randomly concatenated sentence-level data to match the desired input length.

→ We trained the model on the dataset including:

- sentence-level data,
- 1 previous sentence given as a context,
- fixed windows of subword tokens (50, 100, 250, 500), rounded to include full sentences.

```
Netvrdím, že bakteriální celulóza jednou nahradí bavlnu, kůži, nebo jiné látky.
<SEP> Ale myslím, že by to mohl být chytrý a udržitelný přírůstek k našim stále
vzácnějším přírodním zdrojům. <SEP> Možná že se nakonec tyto bakterie neuplatní
v módě, ale jinde. <SEP> Zkuste si třeba představit, že si vypěstujeme lampu,
židli, auto, nebo třeba dům. <SEP> Má otázka tedy zní: Co byste si v budoucnu
nejraději vypěstovali vy?
```

An example document, consisting of 5 sentences separated with a <SEP> tag, before NER tagging and subword encoding.

Quality-Aware Decoding

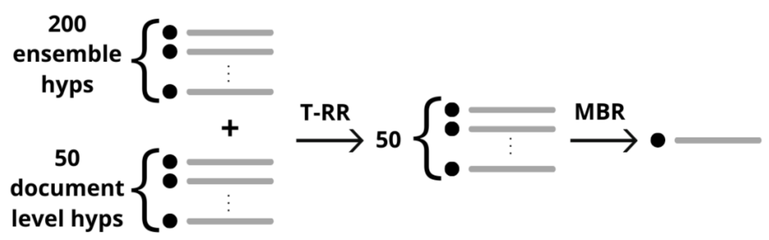
→ We applied a two-stage quality-aware decoding process.

→ First, we tuned a reranker (T-RR), using as features:

- model log-likelihood scores,
- TransQuest QE model trained on DA scores,
- COMET QE model trained on MQM scores,
- COMET QE model trained on DA scores.

→ A total of 250 hypotheses per input sentence (200 from the ensemble and 50 from the document-level model) were reranked by T-RR.

→ Resulting 50 best hypotheses were reranked by using Minimum Bayes Risk (MBR) decoding, using COMET reference-based model as the utility function.



Human Evaluation Results

Ukrainian→Czech

Range	Ave.	Ave. z	System
1	89.6	0.417	HUMAN-A
2-3	85.6	0.182	AMU
2-4	83.5	0.148	HuaweiTSC
4-8	83.5	0.127	Lan-Bridge
3-8	82.0	0.110	CUNI-Transf.
4-8	82.5	0.082	CharlesTranslator
4-8	81.4	0.052	CUNI-JL-JH
4-8	81.9	0.042	Online-B
9-10	80.0	-0.101	Online-A
9-10	77.5	-0.138	Online-G
11	73.9	-0.351	Online-Y
12	69.2	-0.617	ALMAnaCH-Inria

Czech→Ukrainian

Range	Ave.	Ave. z	System
1	85.6	0.295	HUMAN-A
2-5	84.6	0.225	Online-B
2-3	84.1	0.151	AMU
3-6	82.5	0.125	Lan-Bridge
3-6	81.1	0.065	HuaweiTSC
4-8	81.9	0.062	CharlesTranslator
6-8	80.2	0.026	CUNI-JL-JH
6-8	80.2	−0.002	CUNI-Transf.
9-10	79.8	−0.008	Online-G
9-10	79.2	−0.075	Online-A
11	76.0	−0.257	Online-Y
12	68.4	−0.669	ALMAnaCH-Inria