Neural Translation for the EU

Metaforum, Brussels, 8th October 2019

Current coverage of eTranslation

With a few exceptions, all eTranslation MT engines include English as source or target.

Any translation between two non-English languages must use English as pivot

Objective of NTEU: Complement eTranslation's coverage

NTEU will build direct machine translation engines between any of the 24 EU official languages, excluding English

23 x 22 = 506 NMT engines

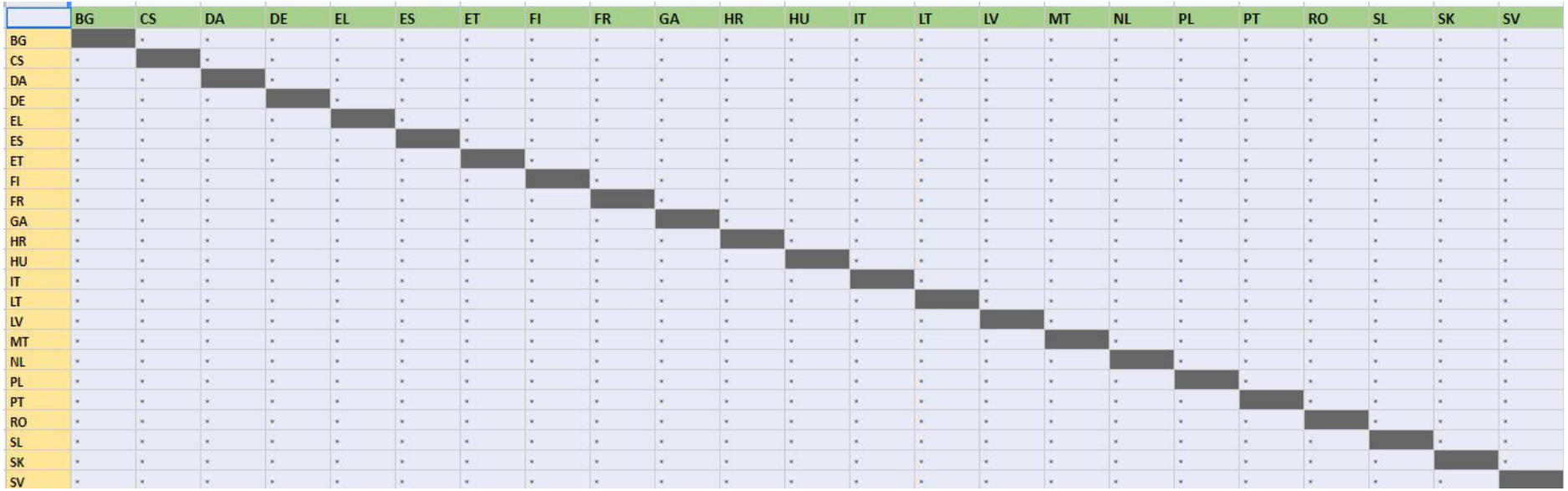
In addition, NTEU will gather and clean data from all language combinations so that engines can be replicated with other technologies in the future.

NTEU outcome

The resulting engines, as well as data and models, shall be made available to the European Commission and to the public administrations of the Member States.

NTEU started in September 2019 and will run until August 2021

Language matrix



	BG	CS	DA	DE	EL	ES	ET	FI	FR	GA	HR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SL	SK	SV
BG																							
CS																							
DA																							
DE																							
EL																							
ES																							
ET																							
FI																							
FR																							
GA																							
HR																							
HU																							
IT																							
LT																							
LV																							
MT																							
NL																							
PL																							
PT																							
RO																							
SL																							
SK																							
SV																							

pangeanic

Spanish, Portuguese, Italian, Dutch, Maltese, Polish, Croatian, French

KantanMT.com

Romanian, German, Danish, Bulgarian, Hungarian, Slovene, Greek, and Irish

TILDE

Latvian, Estonian, Lithuanian, Finnish, Swedish, Czech, Slovak

Common neural network architecture: the Transformer

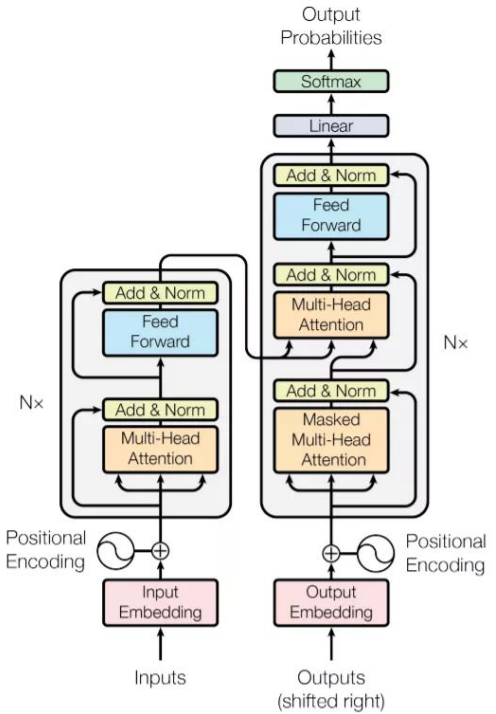


Figure 1: The Transformer - model architecture.

Biggest challenge is to get sufficient training data

NTEU

- 23 * 22 / 2 = 253 parallel corpora
- Minimum estimated 10-15M segments per corpus
- Domain: administration
- Many language pairs are under-resourced
- Some language pairs are severely under-resourced
- Sources: DGT, ELRC, Paracrawl, NEC TM, JRC-Acquis, EMEA-Med, EuBookshop, Europarl, etc. plus copyright-free own resources.

Techniques for under-resourced pairs

- Generation of synthetic data to supplement original data (triangulation, back-translation)
- Transfer learning: cross-language word embeddings, zero-shot translation
- Unsupervised learning on monolingual corpus (e.g. Artetxe, 2019)

Evaluation. Automatic evaluation

- Separate Quality Group within the consortium lead by SEAD, who acts as an independent body, not linked to production (data gathering, engine training)
- Automatic benchmarking against state-of-the-art generic translators (e.g. Google, Bing, DeepL)
- Common test datasets to all language pairs created on purpose, using whole documents and isolated from production data.
- Use of suitable metrics following results of WMT19 Metrics Shared Task.
- If possible evaluate on sub-domains of language, relevant to DSIs

Evaluation. Manual evaluation

- Performed by external agents: universities and research centers through a Public Procurement bid.
- Use of a specific platform developed for purpose.
- Ongoing discussion on alternative ways of manual evaluation:
  - system ranking vs fluency and accuracy evaluation,
  - native target speakers with reference in English vs bilingual speakers

Delivery of engines

Engines will be delivered as docker image (one docker per engine) – with common API interface and home page functionality, able to be run as micro-service.

Dockered engines will include all the necessary dependencies to allow for easy and quick deployment in any major Linux distribution.

They may be used by EU Public Administration and delivered through the European Language Grid to authorised users in the Member States.

They may also be accessed through MT-HUB

Delivery of data

All data used in the project, clean and technologically prepared, will be delivered through the ELRC Share platform, available to authorised users.

Thank you!