TRANSLATION CENTRE
FOR THE BODIES OF THE EUROPEAN UNION

NICE for Covid-19 MLIA

Zuzanna Parcheta

CdT-ALS

SUBTITLING

MODIFICATION

TRANSLATION

EDITING

LANGUAGE CONSULTANCY

REVISION

TERMINOLOGY

Main goal

1) Provide Domain-specific MT engines for the Translation Centre’s clients.

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

European Monitoring Centre
for Drugs and Drug Addiction

European Agency
for Safety and Health
at Work

EUIPO
EUROPEAN UNION
INTELLECTUAL PROPERTY OFFICE

ecdc
EUROPEAN CENTRE FOR
DISEASE PREVENTION
AND CONTROL

1) Decrease the translation effort.

1) Create engines that are fully integrated into CdT’s translations management system.

1) Keep maximum confidentiality in the inference process by assuming an adapted, on-premise infrastructure.

MT Domains

Generic English ↔ 22 UE langs

Public Health English → 22 UE langs

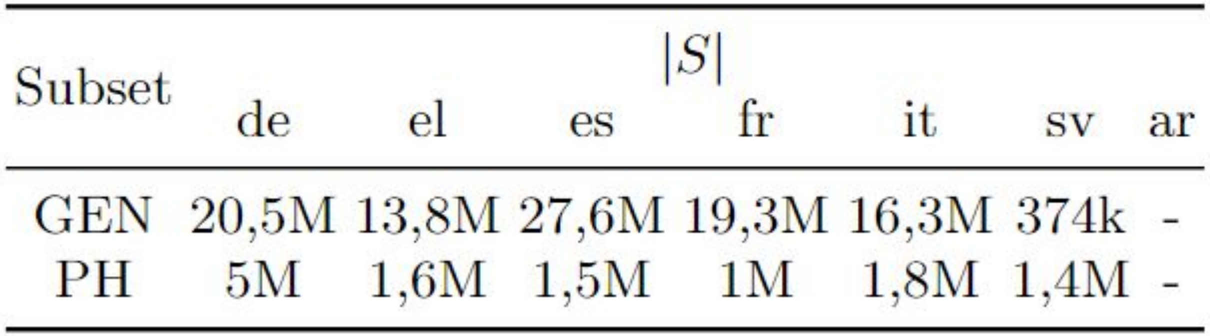
Intellectual Property English ↔ {ES, IT, FR, DE}

Legal English → 22 UE langs

Unconstrained data

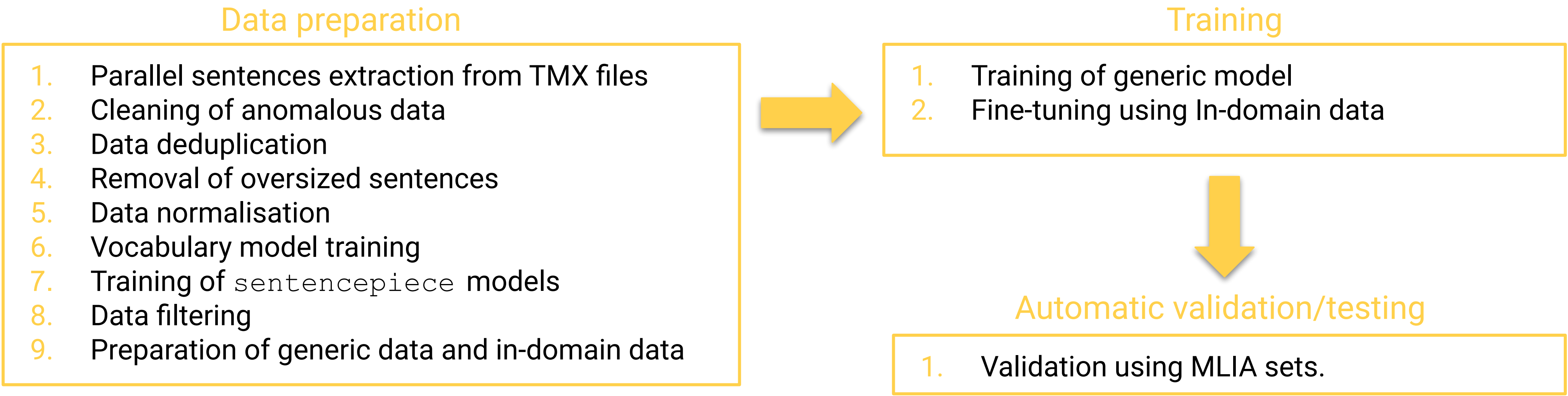
The data is organised depending of quality being 1 the most suitable with the following properties:

- Validated translations from CdT translation memories.
- Non-validated translations from CdT translation memories.
- Verified sentence-based alignments from CdT legacy data.
- Non-CdT data sources (public).
- Synthetic data (CdT and non-CdT)



Subset	$\|S\|$						
	de	el	es	fr	it	sv	ar
GEN	20,5M	13,8M	27,6M	19,3M	16,3M	374k	-
PH	5M	1,6M	1,5M	1M	1,8M	1,4M	-

Implementation



Data preparation

1. Parallel sentences extraction from TMX files
2. Cleaning of anomalous data
3. Data deduplication
4. Removal of oversized sentences
5. Data normalisation
6. Vocabulary model training
7. Training of `sentencepiece` models
8. Data filtering
9. Preparation of generic data and in-domain data

Training

1. Training of generic model
2. Fine-tuning using In-domain data

Automatic validation/testing

1. Validation using MLIA sets.

Implemented systems

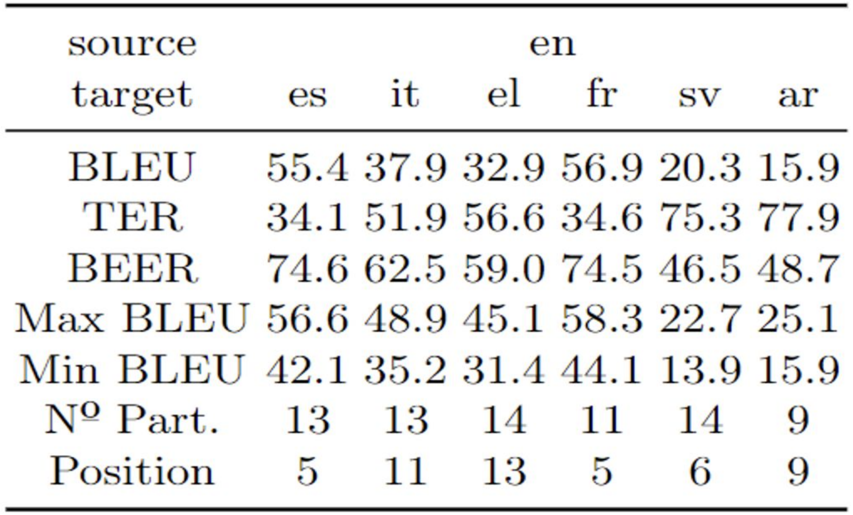
To follow our existing preprocessing pipelines, we included the data provided by organisers into our own data. As the data is external, we included it into the data set of PH domain of quality 4.

3 different systems to generate predictions:

1) Train engine using constrained data (only data from round 2)
2) Generate translation using available Center's engines
3) Train generic engine using constrained+unconstrained GEN data and fine-tune on constrained + PH data

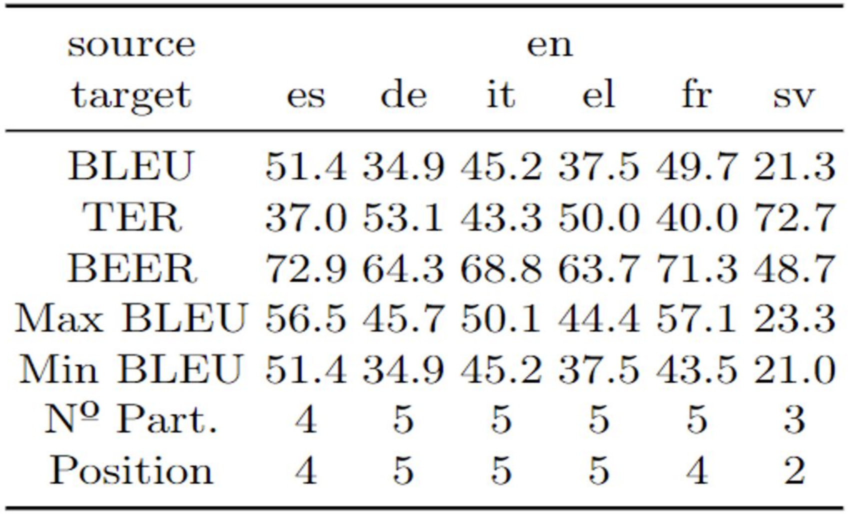
Results

System 1



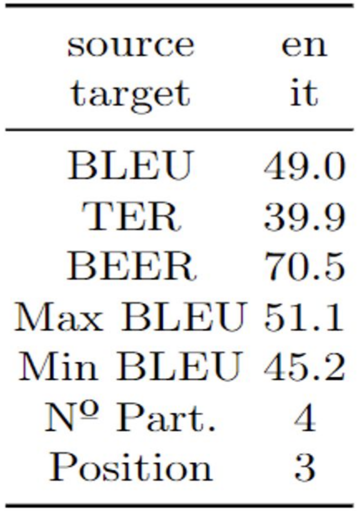
source	en					
target	es	it	el	fr	sv	ar
BLEU	55.4	37.9	32.9	56.9	20.3	15.9
TER	34.1	51.9	56.6	34.6	75.3	77.9
BEER	74.6	62.5	59.0	74.5	46.5	48.7
Max BLEU	56.6	48.9	45.1	58.3	22.7	25.1
Min BLEU	42.1	35.2	31.4	44.1	13.9	15.9
Nº Part.	13	13	14	11	14	9
Position	5	11	13	5	6	9

System 2



source	en					
target	es	de	it	el	fr	sv
BLEU	51.4	34.9	45.2	37.5	49.7	21.3
TER	37.0	53.1	43.3	50.0	40.0	72.7
BEER	72.9	64.3	68.8	63.7	71.3	48.7
Max BLEU	56.5	45.7	50.1	44.4	57.1	23.3
Min BLEU	51.4	34.9	45.2	37.5	43.5	21.0
Nº Part.	4	5	5	5	5	3
Position	4	5	5	5	4	2

System 3*



source	en
target	it
BLEU	49.0
TER	39.9
BEER	70.5
Max BLEU	51.1
Min BLEU	45.2
Nº Part.	4
Position	3

1) Train engine using constrained data (only data from round 2)

2) Generate translation using available Center's engines

3) Train generic engine using constrained+unconstrained GEN data and fine-tune on constrained + PH data

www.cdt.europa.eu

TRANSLATION CENTRE
FOR THE BODIES OF THE EUROPEAN UNION

Thank You

Zuzanna Parcheta

✉ zuzanna.parcheta@cdt.europa.eu

SUBTITLING

MODIFICATION

TRANSLATION

EDITING

LANGUAGE CONSULTANCY

REVISION

TERMINOLOGY