Overview of the "Machine Translation" Task

Francisco Casacuberta[1], Miguel Domingo[1], Mercedes García-Martínez[2], Manuel Herranz[2]

{fcn, midobal}@prhlt.upv.es,
{m.garcia, m.herranz}@pangeanic.com

[1]PRHLT Research Center - Universitat Politècnica de València
[2]Pangeanic / B.I Europa - PangeaMT Technologies Division

Covid-19 MLIA

Virtual Meeting, February 17, 2022

Outline

1. MT task
2. Submissions
3. Results
4. Conclusions

Outline

1. MT task
2. Submissions
3. Results
4. Conclusions

MT task

The goal of the MT task is to generate MT systems focused on Covid-19 related documents for different language pairs.

Examples:

- 30% of children and adults infected with measles can develop complications.
- The MMR vaccine is safe and effective and has very few side effects.

Language pairs

- English–German.
- English–French.
- English–Spanish.
- English–Italian.
- English–Modern Greek.
- English–Swedish.
- English–Arabic. (*New this round*.)

Categories

- **Constrained**: systems which have been trained exclusively with data provided by the organizers (compulsory).
- **Unconstrained**: systems which have been trained using external data and/or resources (optional).

Corpora

From the crawled data we:

- Removed the outliers.
- Selected the best segments for validation and test.
- Balanced the selected segments to have the same representation of each source.

Corpora

		German		French		Spanish		Italian		Modern Greek		Swedish		Arabic	
		En	**De**	**En**	**Fr**	**En**	**Es**	**En**	**It**	**En**	**El**	**En**	**Sv**	**En**	**Ar**
Train	$\lvert S\rvert$	1.5M		2.4M		2.9M		1.0M		674.0K		375.0K		424.4K	
	$\lvert T\rvert$	23.5M	22.1M	45.6M	53.0M	52.4M	60.3M	16.4M	17.2M	11.4M	12.2M	5.5M	5.1M	7.7M	7.5M
	$\lvert V\rvert$	523.9K	847.5K	782.2K	781.4K	850.0K	950.2K	421.2K	501.3K	289.7K	378.7K	180.7K	234.7K	222.2K	360.2K
Validation	$\lvert S\rvert$	4.0K		4.0K		4.0K		4.0K		4.0K		4.0K		4.0K	
	$\lvert T\rvert$	62.2K	61.2K	72.0K	83.9K	72.2K	81.4K	64.6K	69.0K	67.8K	72.5K	56.6K	54.4K	75.9K	74.7K
	$\lvert V\rvert$	13.9K	17.1K	13.2K	14.8K	13.8K	15.8K	14.6K	16.7K	14.0K	18.0K	12.3K	14.1K	16.1K	23.7K
Test	$\lvert S\rvert$	4.0K		4.0K		4.0K		4.0K		4.0K		4.0K		4.0K	
	$\lvert T\rvert$	62.2K	61.0K	72.3K	84.1K	72.2K	81.4K	64.3K	68.7K	67.8K	72.4K	56.5K	54.3K	76.1K	74.5K
	$\lvert V\rvert$	13.8K	17.0K	13.1K	14.8K	13.7K	15.7K	14.4K	16.7K	14.1K	18.2K	12.3K	14.1K	16.2K	23.5K

$|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. M denotes millions and K thousands.

Evaluation

- BLEU.
- TER.
- BEER.

- Approximate Randomization Testing (ART)[1,2].

[1]Riezler, S., Maxwell, J.T.: On some pitfalls in automatic evaluation and significance testing for mt. In: Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 5764 (2005).

[2]`github.com/midobal/covid19mlia-mt-task/blob/master/round2/art`.

Outline

1. MT task
2. Submissions
3. Results
4. Conclusions

Baselines

1. Standard Transformer with only round 2 data.
2. Standard Transformer with rounds 1 and 2 data.

- OpenNMT-py toolkit.
- 32K BPE.

Participants' approaches

- LC:
  - Contrained for all language pairs.
  - Preprocessing: cleaning techniques and inline casing.
  - Language token tag for each source sentence to indicate the target language for multilingual models.
  - Standard transformer in Sockeye instead of Seq2SeqPy used previously (better data loading and support multiple GPUs).
  - Models:
    - Bilingual: better for languages with most data (ES, FR and DE).
    - Multilingual 5 languages excluding Greek and Arabic due to script. IT and SV benefits and even more when oversampling.
    - Multilingual 7 languages. EL and AR benefits and more with oversampling.
  - Contrained data results:
    - Better results for 40K and 50K vocabulary sizes.
    - 1st for ES bilingual.
    - 1st for DE and IT 5 lang multilingual with finetuning.
    - 1st for AR 7 lang multilingual.

- E-Translation:
  - Contrained and unconstrained for 6 lang (all but AR).
  - Cleaning process checking numbers and adding lang identifier.
  - Transformer and big Transformer in MarianNMT.
  - Unconstrained adding TAUS Corona, EMEA and health subset from Euramis corpora.
  - Better architecture is big Transformer 4 model ensambling.
  - Postprocess to normalize puntuation improved by 7 BLEU points for FR in unconstrained.
  - 1st DE unconstrained as 1st round: WMT system with constrained data fine tuning.
  - 1st in all the langs they participated but EL in unconstrained mode.
  - 1st for FR and SV for contrained mode.

- CdT-ASL:
  - Constrained and unconstrained systems.
  - Generic and public health CdT data for unconstrained systems.
  - Cleaning processes
  - Training with big transformer using OpenNMT-tf.
- PROMT:
  - Constrained and unconstrained mode for all langs.
  - Transformer multilingual model with a single encoder and a single decoder with Marian toolkit.
  - Language pairs fine-tuning improves 1-2 additional BLEU points in constrained mode.
  - Unconstrained mode remains as round 1.
  - 1st for constrained EL.

- CUNI-MT:
  - Multilingual models using Transformer in MarianNMT toolkit.
  - They trained jointly on all languages.
  - Constrained mode results are better for transfer learning models.
  - Conclusion:
    - Pretraining a model on a different language pair obtained better results when the corpus size is big.
    - The transfer works also for completely unrelated languages.

Outline

1. MT task
2. Submissions
3. Results
4. Conclusions

English–German

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	5lang-ft-avg	40.3
		ETRANSLATION	ensembleFT	39.9
		LC	5lang-ft	39.8
		ETRANSLATION	ensemble	39.7
		LC	1lang	39.7
	2	PROMT	multilingual-model-round2-tuned-de	39.6
	3	LC	7lang	38.6
	4	PROMT	multilingual-model-round2	39.6
	-	Baseline	Transformer	34.9
		Baseline	Transformer+	34.8
	5	CUNI-MT	transfer	31.8
	6	PROMT	multilingual-model-round1	28.7
	7	CUNI-MT	transfer2	27.5
	8	CUNI-MT	multiling	27.0
Unconstrained	1	ETRANSLATION	wmtFT	45.7
	2	PROMT	Transformer	40.4
	3	ETRANSLATION	singlebigTr	40.0
	4	ETRANSLATION	eTstandardengine	35.4
	-	CdT-ASL*	only-cdt-data	34.9

- 12 different systems from 4 participants.
- Best approaches based on multilingual models and ensembling.
- TER and BEER have similar behavior.

English–French

	Rank	Team	Description	BLEU [↑]
Constrained	1	ETRANSLATION	2	58.3
		ETRANSLATION	1	57.9
		LC	1lang	57.2
		PROMT	multilingual-model-round2-tuned-fr	57.1
	2	CdT-ASL	only-round2-data	56.9
	3	LC	7lang	55.8
		PROMT	multilingual-model-round2	55.4
	-	Baseline	Transformer	54.4
	-	Baseline	Transformer+	53.7
	4	PROMT	multilingual-model-round1	45.4
	5	CUNI-MT	multiling	44.1
Unconstrained	1	PROMT	Transformer	57.1
		ETRANSLATION	generaldenorm	56.9
	2	ETRANSLATION	general	49.9
		CdT-ASL	only-cdt-data	49.7
	3	ETRANSLATION	formal	43.5

- 9 different systems from 5 participants.
- Best approaches based on monolingual models.
- TER presents similar behavior but into fewer clusters.
- BEER behaves similarly.

English–Spanish

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	1lang-avg	56.6
		ETRANSLATION	2	56.1
		ETRANSLATION	1	56.1
		LC	5lang-ft-avg	56.0
	2	CdT-ASL	only-round2-data	55.4
		LC	7lang	55.3
		PROMT	multilingual-model-round2-tuned-es	54.9
	3	PROMT	multilingual-model-round2	53.8
	-	Baseline	Transformer	53.3
	-	Baseline	Transformer+	51.8
	4	CUNI-MT	transfer	48.4
	5	PROMT	multilingual-model-round1	45.1
	6	CUNI-MT	multiling	42.1
Unconstrain.	1	ETRANSLATION	2	56.5
		ETRANSLATION	1	56.0
	2	PROMT	Transformer	53.2
	3	CdT-ASL	only-cdt-data	51.4

- 11 different systems from 5 participants.
- Best approaches based on monolingual and multilingual models.
- TER presents fewer clusters (only one cluster above the baseline).
- BEER behaves similarly.

English–Italian

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC PROMT	5lang-ov-ft-avg multilingual-model-round2-tuned-it	48.9 48.3
	2	LC	5lang-ov	48.0
	3	ETRANSLATION PROMT	4bigTens multilingual-model-round2	47.0 46.8
	4	ETRANSLATION	4bigTensFT	46.7
	5	LC	1lang	45.3
	-	Baseline	Transformer+	43.5
	-	Baseline	Transformer	42.9
	6	CUNI-MT	transfer	38.6
	7	CdT-ASL	only-round2-data	37.9
	8	PROMT	multilingual-model-round1	37.6
	9	CUNI-MT	multiling	35.2
Unconstrained	1	ETRANSLATION	4bigTens	50.1
	2	ETRANSLATION	4bigTensnorm	49.9
	3	CdT-ASL PROMT	round2-data Transformer	49.0 47.8
	4	CdT-ASL	only-cdt-data	45.2

- 11 different systems from 5 participants.
- Best approaches based on multilingual models.
- TER and BEER behave similarly but with fewer clusters.

English–Modern Greek

	Rank	Team	Description	BLEU [↑]
Constrained	1	PROMT	multilingual-model-round2-tuned-el	45.1
	2	LC	7lang-ov-ft-avg	44.7
	3	LC	7lang-ov	44.2
	4	LC PROMT	7lang multilingual-model-round2	43.2 42.1
	5	ETRANSLATION	1	41.7
	6	LC	1lang	41.2
	-	Baseline	Transformer+	39.8
	-	Baseline	Transformer	38.5
	7	ETRANSLATION CUNI-MT	2 transfer	34.9 34.9
	8	CdT-ASL CUNI-MT PROMT	only-round2-data multiling multilingual-model-round1	32.9 32.4 31.4
Unconst.	1	PROMT	Transformer	44.4
	2	ETRANSLATION	2	44.3
	3	ETRANSLATION	1	43.1
	4	CdT-ASL	only-cdt-data	37.5

- 12 different systems from 5 participants.
- Best approaches based on multilingual models.
- TER and BEER behave similarly.

English–Swedish

	Rank	Team	Description	BLEU [↑]
Constrained	1	ETRANSLATION	4bigTens	22.7
		LC	5lang-ov-ft-avg	22.0
		PROMT	multilingual-model-round2-tuned-sv	21.8
		LC	5lang-ov-r2-data	21.8
	2	PROMT	multilingual-model-round2	20.4
	3	CdT-ASL	only-round2-data	20.3
	-	Baseline	Transformer+	19.5
	4	LC	7lang-ov-r1-data	18.3
	5	LC	5lang-r1-data	17.7
	6	PROMT	multilingual-model-round1	17.2
	7	LC	1lang-r1-data	16.7
		Baseline	Transformer	15.3
	8	CUNI-MT	multiling	14.7
		CUNI-MT	transfer	13.9
Unconst.	1	ETRANSLATION	4bigTens	23.3
	2	CdT-ASL	only-cdt-data	21.3
		PROMT	Transformer	21.0

- 12 different systems from 5 participants.
- Best approaches based on multilingual models.
- TER presents fewer clusters.
- BEER behaves similarly.

English–Arabic

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	7lang-ov	25.1
	2	PROMT	multilingual-model-round2-tuned-ar	22.9
	3	LC	7lang	22.0
		PROMT	multilingual-model-round2	21.7
	4	CUNI-MT	transfer	19.1
		LC	1lang	19.1
		Baseline	Transformer	18.8
	5	CUNI-MT	multiling	17.0
	6	CdT-ASL	only-round2-data	15.9
Un.	1	PROMT	Transformer	31.4

- 8 different systems from 4 participants.
- Best approaches based on multilingual models.
- TER presents more clusters.
- BEER behaves similarly.

Outline

1. MT task
2. Submissions
3. Results
4. Conclusions

Conclusions

- 2nd round addressed 7 different language pairs and 2 categories: constrained and unconstrained data usage.
- 5 teams participated in this round.
- Cleaning process improved results.
- Good results for transformer and big transformer architectures.
- Less resource language pairs such as SV, EL and AR benefit from multilingual models.