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, January 12, 2021

Outline

1. MT task
2. Submissions
3. Results
4. Quality assessment
5. Conclusions

Outline

1. MT task
2. Submissions
3. Results
4. Quality assessment
5. 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.

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

		German		French		Spanish		Italian		Modern Greek		Swedish	
		En	**De**	**En**	**Fr**	**En**	**Es**	**En**	**It**	**En**	**El**	**En**	**Sv**
Train	$\|S\|$	926.6K		1.0M		1.0M		900.9K		834.2K		806.9K	
	$\|T\|$	17.3M	16.1M	19.4M	22.6M	19.5M	22.3M	16.7M	18.2M	15.0M	16.4M	14.5M	13.2M
	$\|V\|$	372.2K	581.6K	401.0K	438.9K	404.4K	458.0K	347.7K	416.0K	305.7K	407.5K	298.2K	452.0K
Validation	$\|S\|$	528		728		2.5K		3.7K		3.9K		723	
	$\|T\|$	8.2K	7.6K	17.0K	18.8K	48.9K	56.2K	78.2K	84.0K	73.0K	72.7K	11.4K	10.0K
	$\|V\|$	2.4K	2.6K	4.1K	4.5K	9.7K	10.6K	12.4K	14.9K	10.3K	14.5K	2.6K	2.8K
Test	$\|S\|$	2000		2000		2000		2000		2000		2000	
	$\|T\|$	34.9K	33.2K	33.2K	35.8K	32.6K	34.3K	33.7K	34.2K	42.6K	44.3K	35.3K	30.6K
	$\|V\|$	7.8K	9.6K	6.7K	7.7K	6.7K	7.9K	8.6K	10.4K	9.5K	12.5K	7.1K	8.2K

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

Test selected according to sentence alignments scores.

Evaluation

- BLEU*.
- ChrF.

*Main metric.

- 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] `https://github.com/midobal/mt-scripts/tree/master/art`.

Outline

1. MT task
2. Submissions
3. Results
4. Quality assessment
5. Conclusions

Baselines

1. Standard Transformer.
2. Standard RNN.

- OpenNMT-py toolkit.
- 32K BPE.

Participants' approaches

- PROMT:
  - ALL data concatenated using de-duplication.
  - Language tag added to the source.
  - Transformer MarianNMT.
  - Constrained: 8k sentencepiece.
  - Unconstrained: 16k BPE, Greek no joint BPE (OPUS and WMT).

- CUNI-MT:
  - 1. Online back-translation.
  - 2. Transfer learning with fine-tuning of a low resource child model.
  - 3. Multilingual model of Latin languages.
  - Transformer XLM toolkit, 30k BPE.
- CUNI-MTIR:
  - Transformer MarianNMT, 32k BPE.
  - Unconstrained: UFAL Medical corpus.

- LC:
  - Multilingual, no Greek due to script.
  - Language tag in source, fine tuning.
  - Transformer seq2seqPy, 50k for single and 70k for multilingual sentencepiece.
  - More improvements for German using multilingual, not much for French.
- LIMSI:
  - 32K BPE.
  - Transformer BERT fairseq toolkit.

- TARJAMA-AI:
  - Single models trained with all the language pairs.
  - Language token for other languages, desired language is kept untagged.
  - Oversamble desired target language.
- E-Translation:
  - Transformer MarianNMT.
  - German constrained: Transfer learning, 12K SentencePiece.
  - German unconstrained: WMT system and fine tuning with constrained data.
  - French constrained: small and big (missing info).
  - French unconstrained: gen, phwt and eufl (missing info).
- ACCENTURE (missing report):
  - Multilingual BART model.

Outline

1. MT task
2. Submissions
3. Results
4. Quality assessment
5. Conclusions

Constrained English–German

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	CUNI-MT	transfer2	31.6	0.600
	CUNI-MT	base	31.4	0.596
	CUNI-MT	transfer1	31.3	0.595
	PROMT	multilingual	31.1	0.599
2	ETRANSLATION	basetr	30.4	0.593
3	CUNI-MT	transfer2	29.8	0.584
	LC	multilingual	29.5	0.584
4	Baseline	transformer	28.1	0.573
	LC	transformer	26.7	0.556
5	TARJAMA-AI	base3	25.6	0.564
6	TARJAMA-AI	base2	25.0	0.559
7	CUNI-MTIR	r1	19.7	0.494
8	Baseline	RNN	17.9	0.479
	TARJAMA-AI	base	17.7	0.488

- 12 different systems from 6 participants.
- Best approaches were based on transfer learning, standard NMT with back-translation and multilingual NMT.

Constrained English–French

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	multilingual	49.6	0.711
2	ETRANSLATION	small	49.1	0.707
	LC	multilingual	49.0	0.705
	LC	transformer	48.9	0.703
	CUNI-MT	base	48.4	0.703
	CUNI-MT	multiling	48.0	0.700
	ETRANSLATION	big	47.4	0.695
	Baseline	transformer	47.3	0.693
	CUNI-MT	transfer2	47.1	0.693
3	LIMSI	trans	43.5	0.660
4	CUNI-MTIR	r1	34.9	0.605
-	Baseline	RNN	34.3	0.596
5	TARJAMA-AI	base	26.8	0.567
6	ACCENTURE	mbart	15.8	0.464

- 12 different systems from 8 participants.
- Best approaches were based on multilingual NMT, transfer learning and standard NMT with back-translation.

Constrained English–Spanish

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	multilingual	48.3	0.702
2	CUNI-MT	transfer1	47.9	0.699
	CUNI-MT	transfer2	47.6	0.698
	LC	multilingual	47.5	0.695
	Baseline	transformer	47.4	0.694
	CUNI-MT	multiling	47.3	0.692
	CUNI-MT	base	47.3	0.691
-	Baseline	RNN	35.6	0.609
3	CUNI-MTIR	r1	32.9	0.591
4	TARJAMA-AI	base	30.9	0.593
5	ACCENTURE	mbart	17.4	0.474

- 9 different systems from 6 participants.
- Best approaches were based on multilingual NMT, transfer learning and standard NMT with back-translation.

Constrained English–Italian

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	multilingual	29.6	0.585
2	LC	multilingual	28.4	0.572
	CUNI-MT	transfer2	28.3	0.574
	CUNI-MT	multiling	28.3	0.574
-	Baseline	transformer	26.9	0.560
3	TARJAMA-AI	base	19.2	0.494
-	Baseline	RNN	17.0	0.473

- 5 different systems from 4 participants.
- Best approaches were based on multilingual NMT and transfer learning.

Constrained English–Modern Greek

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	multilingual	27.2	0.523
2	CUNI-MT	transfer1	24.7	0.496
3	CUNI-MT	base	24.1	0.484
	Baseline	transformer	22.6	0.471
-	Baseline	RNN	12.8	0.365

- 3 different systems from 2 participants.
- Best approaches were based on multilingual NMT and transfer learning.

Constrained English–Swedish

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	multilingual	30.7	0.595
	LC	multilingual	30.4	0.589
	CUNI-MT	transfer2	30.1	0.590
2	CUNI-MT	transfer	28.5	0.578
-	Baseline	transformer	27.8	0.566
3	CUNI-MT	base	26.6	0.561
4	CUNI-MTIR	r1	25.1	0.541
-	Baseline	RNN	19.2	0.481
5	TARJAMA-AI	base	11.2	0.443

- 7 different systems from 5 participants.
- Best approaches were based on multilingual NMT and transfer learning.

Unconstrained English–German

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	ETRANSLATION	wmtfinetune	44.4	0.686
2	ETRANSLATION	wmt	44.1	0.683
3	PROMT	transformer	41.2	0.666
4	CUNI-MTIR	r1	20.0	0.499

- 4 different systems from 3 participants.
- Best approaches were based on a WMT system and multilingual NMT.

Unconstrained English–French

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	59.5	0.767
2	ETRANSLATION	gen	52.9	0.742
3	LIMSI	indom	51.2	0.721
4	ETRANSLATION	phwt	50.1	0.724
	LIMSI	trans	49.3	0.710
	LIMSI	bert	49.3	0.703
	LIMSI	mlia	48.5	0.705
5	ETRANSLATION	eufl	47.9	0.712
6	CUNI-MTIR	r1	33.0	0.590

- 8 different systems from 4 participants.
- Best approaches were based on multilingual NMT and Transformer enriched with external data from OPUS and WMT.

Unconstrained English–Spanish

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	58.2	0.762
2	CUNI-MTIR	r1	32.1	0.582

- 2 different systems from 2 participants.
- Best approaches were based on multilingual NMT and Transformer enriched with WMT and OPUS (PROMT) and in-domain data (CUNI-MTIR).

Unconstrained English–Italian

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	38.0	0.642

- A single system.
- Based on multilingual NMT enriched with OPUS and WMT data.

Unconstrained English–Modern Greek

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	42.4	0.652

- A single system.
- Based on multilingual NMT enriched with OPUS and WMT data.

Unconstrained English–Swedish

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	41.3	0.671
2	CUNI-MTIR	r1	24.0	0.514

- 2 different systems from 2 participants.
- Best approaches were based on multilingual NMT and Transformer enriched with WMT and OPUS (PROMT) and in-domain data (CUNI-MTIR).

Outline

1. MT task
2. Submissions
3. Results
4. Quality assessment
5. Conclusions

Quality assessment

Team	Description	Reference		Post-edition	
		BLEU [↑]	chrF [↑]	BLEU [↑]	chrF [↑]
PROMT	multilingual	45.1	0.682	43.9	0.672
CUNI-MT	transfer1	46.2	0.686	43.8	0.672
CUNI-MT	transfer2	46.0	0.686	43.4	0.671
LC	multilingual	45.8	0.684	43.5	0.669
Baseline	transformer	45.4	0.682	43.9	0.670
CUNI-MT	multiling	44.7	0.677	43.0	0.664
CUNI-MT	base	45.0	0.675	42.4	0.660
Baseline	RNN	34.6	0.603	32.3	0.589
CUNI-MTIR	r1	31.4	0.583	30.8	0.578
TARJAMA-AI	base	29.2	0.583	26.9	0.569
ACCENTURE	mbart	16.7	0.466	16.0	0.460

- 500 Spanish segments post-edited by a team of professional translators.
- 18.8 TER from the reference and its post-edited version.
- Evaluation is consistent using the reference and its post-edited version.

Outline

1. MT task
2. Submissions
3. Results
4. Quality assessment
5. Conclusions

Conclusions

- This first round addressed 6 different language pairs and was divided into two categories: *constrained* and *unconstrained*.
- 8 different teams took part in this round.
- The most successful approaches were based on multilingual MT and transfer learning.
- PROMT's approach:
  - Best results for all language pairs in both categories except for unconstrained English–German and constrained English–German sharing the first position with CUNI-MT.
  - Multilingual system trained using all language pairs, a smaller vocabulary and sentencepiece.
- In general, the difference between systems from one rank and the next one is small. The RNN baseline delimits the point in which there is a significant drop of translation quality between ranks.