



# Machine Translation at NAVER and NAVER LABS Europe

ACL 2020  
30 min Interactive Session  
2020-07-08


# Agenda

- Introduction to Papago MT and NAVER LABS Europe MT
- Research topics
  - COVID19 model release
  - Honorific translation
  - Robustness in MT (WMT19)
  - Robustness in MT (recent)
  - Evaluation
- Q&A

# Machine translation at **NAVER LABS** Europe

<b>NMT in the context of NAVER LABS Europe</b>	<b>Challenge of scalability</b>
<ul style="list-style-type: none"><li>• AI research centre located in the French Alps</li><li>• 100 scientists organized around competencies in NLP, Computer Vision, Machine Learning &amp; Optimization, Search &amp; Recommendation, UX &amp; ethnography</li><li>• The NMT project combines these competencies to solve language related problems</li></ul>	<p>Handling a massive number of:</p> <ul style="list-style-type: none"><li>• Users</li><li>• Size of documents</li><li>• Languages</li><li>• Domains</li></ul> <p>is currently still challenge for current NMT technologies.</p>
<b>Challenge of multimodality</b>	<b>Challenge of controllability</b>
<p>Current MT systems focus on text as input but other modalities such as speech and images are more and more prevalent</p>	<p>Our current solutions for fine grained control of NLG/NMT models, or for avoiding catastrophic failures, are for now limited</p>

# Machine translation at **NAVER**

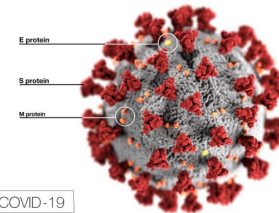
 To be covered today!

MT modeling	Multilingual NLP
<ul style="list-style-type: none"><li>• Sequence-to-sequence modeling</li><li>• Aims for best MT quality for Chinese/Japanese/Korea &amp; English</li><li>• Aims for best MT quality for K-pop, etc.</li><li>• Broader contexts (text, image, speech), <b>controllable models</b> (honorifics, diversity), multilingual models, speech enhanced translation, <b>evaluation</b></li></ul>	<ul style="list-style-type: none"><li>• Multilingual text classification</li><li>• Multilingual sequence labeling</li><li>• Multilingual language modeling</li><li>• Multilingual sentence similarity, quality estimation</li></ul>
ML Engineering	Papago
<ul style="list-style-type: none"><li>• Model compression</li><li>• Inference throughput/latency optimization (ex: Non-autoregressive decoding)</li><li>• Data, training, deployment pipelines for scalability</li></ul>	<ul style="list-style-type: none"><li>• <a href="https://papago.naver.com">https://papago.naver.com</a></li><li>• Most popular translation service in Korea</li><li>• Text/Image/Voice/Website/Offline translation</li><li>• <a href="#">Papago Gym</a> (User participation)</li></ul>

## COVID-19 translation model

Vassilina Nikoulina (NAVER LABS Europe)

# Context: Covid-19 crisis



**NAVER LABS Europe:** Alexandre Berard, Vassilina Nikoulina, Matthias Galle @naverlabs. COVID-19

**Papago:** Zae Myung Kim, Lucy Park @navercorp.com

## Objective

- Creation of big **multilingual** and **multi-domain** translation model

## Potential applications

- Assist human translations in translating Covid-19-related documents from French, Spanish, Italian, German or Korean into English
- Enabling large-scale multilingual content analysis of the documents related to Covid19 pandemics
  - e.g. User-generated contents, governmental guidelines, other?
- other?

# Approach

- Gathering training data for different languages and domains
  - Languages covered (*most touched countries at the moment of creation of the model*):  
*French, Spanish, Italian, German, Korean → English*
- Biomedical data is available for some language pairs (English, French-English, German-English), but very scarce or absent for others (English, Italian-English, Korean-English)  
→ we train multi-domain model which enables zero-shot domain transfer
- Creation of biomedical test-sets
  - Gathering existing datasets (French, Spanish, German)
  - Creating datasets: Korean
- Adapting parameters of transformer-big model based on previous experiments:
  - *Transformer.big* architecture used as a basis
  - Extending encoder capacity to better handle multiple languages
  - Decreasing decoder capacity to keep model size reasonable

## Some results

Language	Model	News	Medline	IWSLT
French	Ours	<b>41.00</b>	<b>36.16</b>	<b>41.09</b>
	SOTA	40.22*	35.56 <sup>‡</sup>	–
	OPUS-MT	36.80	33.60	38.90
German	Ours	<b>41.28</b>	<b>29.76</b>	31.55
	SOTA	40.98 <sup>†</sup>	28.82 <sup>‡</sup>	<b>32.01<sup>†</sup></b>
	OPUS-MT	39.50	28.10	30.30
Spanish	Ours	<b>36.63</b>	<b>46.18</b>	<b>48.79</b>
	SOTA	–	43.03 <sup>‡</sup>	–
	OPUS-MT	30.30	43.30	46.10
Italian	Ours			<b>42.18</b>
	SOTA			–
	OPUS-MT			39.70
Korean	Ours			<b>21.33</b>
	SOTA			–
	OPUS-MT			17.60

## If you want to know more:

- Blog post : <https://europe.naverlabs.com/blog/a-machine-translation-model-for-covid-19-research/>
- NLP-Covid workshop submission: [https://openreview.net/forum?id=2\\_c3GLAEIQL](https://openreview.net/forum?id=2_c3GLAEIQL)
- To play with the model: <https://github.com/naver/covid19-nmt>



# Honorific translation

Kweonwoo Jung (NAVER)

## Motivation

- Provide culturally adequate translation results
  - “Hi” to Elderly vs “Hi” to Peers is different in Korean
- when the honorific is not aligned to the context, it can be RUDE..

## Contribution

- provide “Honorific” option in [English > Korean] translation
- Papago users were guaranteed to be polite

Korean ▾ Honorific ☐

안녕

annyeong

Korean ▾ Honorific ☒

안녕하세요

annyeonghaseyo

## Approach

- Two possible directions
  - A : post-edit Korean output into honorific text
  - B : use NMT model with honorific tag to generate honorific text
- Option B is selected, since source context helps generate better honorific text
- One-liner : honorific tag based NMT training
  - Source-side or target-side?
  - Which position?
  - as a token? or as an embedding (like positional embedding)?
- General Process
  - Given a bilingual text [XX -- Korean], tag Korean with either honorific or not
  - Append pairs with honorific Korean to original bilingual corpus
    - want to avoid baseline model becoming non-honorific
  - Train NMT
  - Make inference with controllable honorific tag

## Results

- No negative effect on Baseline model
- Controllable honorific translation

## Challenges

- Error propagation from honorific Tagger
- Coverage of Honorifics
  - High coverage in verb, especially stem + ending word
  - Low coverage in noun, pronoun
- Degree of Honorifics
  - Politeness to your older brother vs Politeness to your professor vs Politeness to your King etc

## Robustness in MT (WMT19)

Ioan Calapodescu (NAVER LABS Europe)

# NMT for User Generated Content (UGC)

## Motivation

- Off-the-shelf models have problems translating UGC: blogs (like Reddit or Naver Cafe), comments and reviews (like Naver Maps or Google Maps), social media (like Twitter).
- Part of the problem is due to the noise in the input and it highlights the lack of robustness of our models

# WMT 2019 Robustness Shared Task

We participated to the 1st WMT Shared Task on Robustness: Translation Reddit comments in FR/EN/JP

## Contributions

- Data filtering techniques: bad training data is part of the problem
- Robustness tricks: natural noise generation, inline casing and preprocessing (emojis)
- Domain adaptation: noisy data could be considered as a specific domain

Bérard, Alexandre, Ioan Calapodescu, and Claude Roux. "Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task."

Bérard, Alexandre, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. "Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness." 3S. All rights reserved.

System	FR→EN	EN→FR	JA→EN	EN→JA
Baseline	25.6	22.1	5.8	8.4
Transformer	40.9	37.0	-	-
Transformer + MTNT + tags	45.0	39.0	13.7	16.4
NLE single	47.0	41.0	14.7	16.5
<b>NLE ensemble</b>	<b>47.9 (1)</b>	<b>41.4 (1)</b>	<b>16.4 (1)</b>	<b>17.7 (1)</b>
NTT	-	-	14.8 (2)	16.9 (1)
Baidu + OSU	43.6 (3)	36.4 (3)	-	-
CUNI	44.8 (2)	38.5 (2)	-	-
JHU	40.2 (4)	-	12.0 (3)	14.7 (3)

System	FR→EN	EN→FR	JA→EN	EN→JA
<b>NLE ensemble</b>	<b>85.3 (1)</b>	<b>75.5 (1)</b>	<b>74.1 (1)</b>	<b>63.9 (2)</b>
NTT	-	-	71.3 (2)	<b>66.5 (1)</b>
Baidu + OSU	80.6 (3)	71.5 (2)	-	-
CUNI	82.0 (2)	66.3 (3)	-	-
JHU	76.3 (4)	-	65.4 (3)	58.5 (3)

## Robustness in MT

Stephane Clinchant (NAVER LABS Europe)



# Robustness in MT Models

- What do we mean by robustness in MT (and in ML) ?
- How can we measure it ?
  - Metrics  $\rightarrow \Delta$  Metrics , ...
  - Noisy Test Sets
- How can we use prior knowledge ? (ex: BERT)
  - [On the use of BERT for Neural Machine Translation](#) EMNLP'19 WNGT
- Can we go beyond data augmentation ?
  - Robust models by Design (e.g Adversarial Networks)
  - Ongoing Work: A simpler alternative to existing approaches

# Evaluation

Jihyung Moon (NAVER)

<https://arxiv.org/abs/2004.13937>

# MT Evaluation Methods

## Motivation

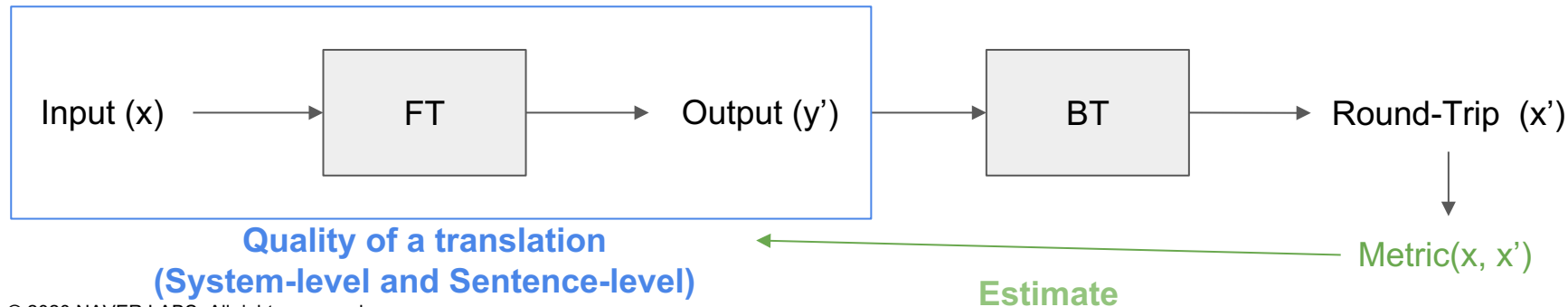
Evaluation Method	Description	
Human Evaluation	Translation output is evaluated by (bilingual) <b>human</b>	
Reference-based Metric	Automatic similarity measurement between translation output and <b>reference (human-generated golden-truth)</b> e.g., BLEU, chrF, ....	
Quality Estimation	Automatic similarity measurement between translation output and <b>input</b>	

Accurate

Cheap

# Round-Trip Translation based QE Metric

- Round-Trip Translation (RTT)
  - Input (x)  $\rightarrow$  Forward Translation (FT)  $\rightarrow$  Output (y')  $\rightarrow$  Backward Translation (BT)  $\rightarrow$  **Round-Trip sentence (x')**
- RTT-based QE Metric
  - **Metric(x, x')** is a scalar function computing the similarity of x and x'
    - **Examples of Metric:**  
BLEU, chrF, METEOR, **BERTScore**, SentBERT cosine similarity ...



# Revisiting Round-Trip Translation based QE Metric

- **Lexical-level** metric vs. **Semantic-level** metric
  - BLEU, chrF vs. BERTScore, SentBERT cosine similarity

Input (en)	'We know it won't change students' behaviour instantly.
Reference (de)	Wir wissen, dass es das Verhalten der Studenten nicht sofort ändern wird.
Output (de)	„Wir wissen, dass es das Verhalten der Schüler nicht sofort ändern wird.
Round-trip (en)	“We know that it will not change student behavior immediately.
RTT-SENTBLEU: 14.99 (rank: 1947/1997)	
RTT-SBERT(*): 98.07 (rank: 1001/1997)	
RTT-BERTSCORE(*): 97.04 (rank: 1033/1997)	

- Recently, RTT is used to generate paraphrases

- Lexical-level metrics (e.g., BLEU, chrF) are failed to measure paraphrases

- **What if we use semantic-level metrics?**

# Which BT system should we use?

- **Online system** > WMT trained system
  - Training set of an online system: not constrained to WMT news corpus (out-of-domain)
  - Training set of WMT systems: constrained to WMT news corpus
  - WMT en-de experiments

Backward translations		Pearson correlations				Variance ( $\times 10^{-4}$ )	
Systems	BLEU	RTT-BLEU	RTT-CHRf	RTT-SBERT	RTT-BERTSCORE	RTT-SBERT	RTT-BERTSCORE
Google	<b>46.96</b>	0.797	0.853	0.941	0.951	5.08	1.96
Microsoft	42.68	<b>0.845</b>	<b>0.877</b>	<b>0.948</b>	0.955	5.12	2.07
Amazon	40.89	0.776	0.804	0.941	<b>0.956</b>	4.86	1.88
Facebook-FAIR	42.17	0.788	0.865	0.940	0.934	4.84	1.27
Transformer Big (100k)	38.96	0.739	0.818	0.939	0.937	4.58	1.57
Transformer Big (40k)	36.38	0.707	0.795	0.938	0.935	4.22	1.36
Transformer Big (20k)	34.75	0.617	0.759	0.931	0.860	3.97	1.15
Transformer Big (10k)	31.30	0.509	0.749	0.908	0.789	3.17	0.91

Even with similar BLEU score, metrics are more successful when using the online system

# Sensitivity to BT system

- **RTT-SBERT, RTT-BERTScore** > RTT-BLEU, RTT-chrF
  - In terms of
    - 1) Pearson correlation
    - 2) Robustness toward BT system
  - WMT en-de experiments

Backward translations		Pearson correlations				Variance ( $\times 10^{-4}$ )	
Systems	BLEU	RTT-BLEU	RTT-CHRf	RTT-SBERT	RTT-BERTScore	RTT-SBERT	RTT-BERTScore
Google	<b>46.96</b>	0.797	0.853	0.941	0.951	5.08	1.96
Microsoft	42.68	<b>0.845</b>	<b>0.877</b>	<b>0.948</b>	0.955	5.12	2.07
Amazon	40.89	0.776	0.804	0.941	<b>0.956</b>	4.86	1.88
Facebook-FAIR	42.17	0.788	0.865	0.940	0.934	4.84	1.27
Transformer Big (100k)	38.96	0.739	0.818	0.939	0.937	4.58	1.57
Transformer Big (40k)	36.38	0.707	0.795	0.938	0.935	4.22	1.36
Transformer Big (20k)	34.75	0.617	0.759	0.931	0.860	3.97	1.15
Transformer Big (10k)	31.30	0.509	0.749	0.908	0.789	3.17	0.91

**Semantic-level metrics outperform the other metrics and are robust to the type and performance of the BT systems**

# Performance across Language Pairs

BT system = Google  
(*∴ supported language pairs and performance*)

- System-level performance
  - BLEU, chrF > RTT-SBERT, RTT-BERTScore > RTT-BLEU, RTT-chrF

src lang tgt lang	de en	fi en	gu en	kk en	lt en	ru en	zh en	avg. (std.)	en cs	en de	en fi	en gu	en kk	en lt	en ru	en zh	avg. (std.)
n	16	12	11	11	11	14	15		11	22	12	11	11	12	12	12	
BLEU*	.849	.982	.834	.946	.961	.879	.899	.907 (.057)	.897	.921	.969	.737	.852	.989	.986	.901	.907 (.084)
CHRf*	.917	.992	.955	.978	.940	.945	.956	.955 (.025)	.990	.979	.986	.841	.972	.981	.943	.880	.947 (.056)
SACREBLEU-BLEU*	.813	.985	.834	.946	.955	.873	.903	.901 (.065)	.994	.969	.966	.736	.852	.986	.977	.801	.910 (.100)
SACREBLEU-CHRf*	.910	.990	.952	.969	.935	.919	.955	.947 (.028)	.983	.976	.980	.841	.967	.966	.985	.796	.937 (.074)
QE as a Metric																	
Individual Best*	.850	.930	.566	.324	.487	.808	.947	- (-)	.871	.936	.907	.314	.339	.810	.919	.118	- (-)
YiSi-2*	<b>.796</b>	.642	.566	.324	.442	.339	<b>.940</b>	.578 (.232)	.324	.924	.696	.314	.339	.055	.766	.097	.439 (.319)
RTT-BLEU	.130	<b>.827</b>	.641	.859	.596	.295	.825	.596 (.284)	-.625	.797	.417	.608	.930	-.334	.572	-.599	.221 (.637)
RTT-CHRf	.495	.810	<b>.778</b>	.776	.692	.524	.875	.707 (.146)	-.408	.842	.487	.586	.423	-.153	.750	-.310	.277 (.493)
RTT-SBERT	.761	-	-	-	-	<b>.867</b>	.889	.839 (.005)	.470	.941	.804	.710	.950	<b>.410</b>	.833	<b>.256</b>	<b>.672</b> (.261)
RTT-BERTScore	.654	.819	.729	<b>.889</b>	<b>.712</b>	.816	.912	<b>.790</b> (.095)	<b>.473</b>	<b>.951</b>	<b>.819</b>	<b>.737</b>	<b>.966</b>	.342	<b>.869</b>	.071	.654 (.324)

**Table 3:** Pearson correlations of system-level metrics with human judgments on WMT19. The best correlations of QE-as-a-metric within the same language pair are highlighted in bold. \* denotes that reported correlations are from WMT19 metrics task (Ma et al., 2019).



# Performance across Language Pairs

BT system = Google  
(*∵ supported language pairs and performance*)

- Sentence-level performance
  - BLEU, chrF > RTT-SBERT, RTT-BERTScore > RTT-BLEU, RTT-chrF

src lang tgt lang	de en	fi en	gu en	kk en	lt en	ru en	zh en	avg. (std.)	en cs	en de	en fi	en gu	en kk	en lt	en ru	en zh	avg. (std.)
n	85k	38k	31k	27k	22k	46k	31k		27k	100k	32k	11k	18k	17k	24k	19k	
SENTBLEU*	.056	.233	.188	.377	.262	.125	.323	.223 (.111)	.367	.248	.396	.465	.392	.334	.469	.270	.368 (.081)
CHRF*	.122	.286	.256	.389	.301	.180	.371	.272 (.096)	.455	.326	.514	.534	.479	.446	.539	.301	.449 (.091)
QE as a Metric																	
Individual Best*	.022	.211	-.001	.096	.075	.089	.253	- (-)	.069	.236	.351	.147	.187	.003	.226	.044	- (-)
YiSi-2*	<b>.068</b>	.126	-.001	.096	.075	<b>.053</b>	.253	.096 (.080)	<b>.069</b>	<b>.212</b>	.239	.147	.187	.003	-.155	.044	.093 (.131)
RTT-SENTBLEU	-.169	.095	.111	.140	.086	-.104	.168	.047 (.130)	-.122	-.001	.088	.374	.399	-.110	.157	-.106	.085 (.211)
RTT-CHRF	-.114	.141	<b>.184</b>	.130	.099	-.050	.195	.083 (.119)	-.093	.055	.119	.395	.310	-.069	.195	-.075	.105 (.185)
RTT-SBERT	-.066	-	-	-	-	-.013	.225	.049 (.024)	.025	.169	.268	.444	.503	<b>.070</b>	.371	<b>.064</b>	.239 (.185)
RTT-BERTSCORE	-.085	<b>.185</b>	.167	<b>.204</b>	<b>.118</b>	-.020	<b>.255</b>	<b>.118</b> (.125)	.065	.194	<b>.292</b>	<b>.494</b>	<b>.579</b>	.069	<b>.391</b>	.056	<b>.268</b> (.205)

**Table 4:** Kendall's  $\tau$  formulation of segment-level metric scores with human judgments on WMT19. The best correlations of QE-as-a-metric within the same language pair are highlighted in bold. For some language pairs, QE metrics obtain negative correlations. \* denotes that reported correlations are from WMT19 metrics task (Ma et al., 2019).

# Sensitivity to FT system

BT system = Google  
(*∵ supported language pairs  
and performance*)

- SMT vs. NMT
  - SMT: WMT12 submissions
  - NMT: WMT19 submissions
  - **RTT-SBERT and RTT-BERTScore demonstrate the most promising performance regardless of the FT systems.**

Language Pairs	Systems (n)	Pearson correlations				
		BLEU	RTT-BLEU	RTT-CHRF	RTT-SBERT	RTT-BERTSCORE
English–Czech	SMT (12)	0.615	0.261	0.342	0.482	<b>0.620</b>
	NMT (11)	0.897	-0.625	-0.408	0.470	<b>0.473</b>
English–German	SMT (12)	0.582	0.523	0.553	0.742	<b>0.765</b>
	NMT (22)	0.921	0.797	0.842	0.941	<b>0.951</b>
German–English	SMT (13)	0.841	0.530	0.374	<b>0.712</b>	0.682
	NMT (16)	0.849	0.130	0.495	<b>0.761</b>	0.654

# Conclusions

- We reconsider RTT with suitable semantic-level metrics, specifically SBERT and BERTScore in our settings, and show it can be used to measure translation quality.
- We observe RTT methods using SBERT and BERTScore are robust to the choice of BT systems.
- We present RTT with semantic similarity measurements consistently exhibit high-performance across different FT systems: SMT and NMT.
- We find the paraphrase detection ability of metrics is related to the performance of RTT-based QE.

## Q&A

- NAVER: [dl\\_papago\\_mt\\_recruit@navercorp.com](mailto:dl_papago_mt_recruit@navercorp.com)
- NAVER LABS Europe: [europe.naverlabs.com](https://europe.naverlabs.com)