Quality Estimation and Automatic Post-editing in the Neural Machine Translation Era

Lucia Specia

Imperial College/University of Sheffield
l.specia@sheffield.ac.uk

HAT Workshop, Dublin, August 19th 2019

Outline

Outline

Neural Machine Translation

- From **bridging human gap**

Neural Machine Translation

- From **bridging human gap**

Google AI Blog

The latest news from Google AI

A Neural Network for Machine Translation, at Production Scale

Tuesday, September 27, 2016

Posted by Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team

Ten years ago, we announced the launch of Google Translate, together with the use of Phrase-Based Machine Translation as the key algorithm behind this service. Since then, rapid advances in machine intelligence have improved our speech recognition and image recognition capabilities, but improving machine translation remains a challenging goal.

Today we announce the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for machine translation quality. Our full research results are described in a new technical report we are releasing today: "*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*" [1].

A few years ago we started using Recurrent Neural Networks (RNNs) to directly learn the mapping between an input sequence (e.g. a sentence in one language) to an output sequence (that same sentence in another language) [2]. Whereas Phrase-Based Machine Translation (PBMT) breaks an

Neural Machine Translation

- To **human parity**

Neural Machine Translation

- To **human parity**

Microsoft | Research Research areas Researcher tools Programs & Events Careers People Blogs & Podcasts Labs & Locations A

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou

Download BibTex

March 2018
arXiv:1803.05567
View Publication

Machine translation has made rapid advances in recent years. Millions of people are using it today in online translation systems and mobile applications in order to communicate across language barriers. The question naturally arises whether such systems can approach or achieve parity with human translations. In this paper, we first address the problem of how to define and accurately measure human parity in translation. We then describe Microsoft's machine translation system and measure the quality of its translations on the widely used WMT 2017 news translation task from Chinese to English. We find that our latest neural machine translation system has reached a new state-of-the-art, and that the translation quality is at human parity when compared to professional human translations. We also find that it significantly exceeds the quality of crowd-sourced non-professional translations.

View Publication

Groups

Machine Translation

Research Areas

Artificial intelligence

Neural Machine Translation

- To **superhuman** performance

Neural Machine Translation

- To **superhuman** performance

Facebook AI leads in 2019 WMT international machine translation competition

August 01, 2019 Written by Nathan Ng, Sergey Edunov, Michael Auli

With hundreds of languages used by people on our platforms and thousands more spoken around the world, developing powerful and flexible machine translation systems has long been a research focus for Facebook. Today we are proud to announce that Facebook AI models achieved first place in several language tasks included in this year's annual news translation competition, hosted by the Fourth Conference on Machine Translation (also known as WMT). Our models outperformed all other entrants' models in the four tasks we participated in, including English to German, the most competitive task in the contest, with entries drawn from a wide range of high-performing research teams. For this language direction, our translations have been declared superhuman by the WMT organizers, meaning that human evaluators preferred them over translations done by human experts.

Neural Machine Translation

- To **superhuman** performance

Facebook AI leads in 2019 WMT international machine translation competition

August 01, 2019 Written by Nathan Ng, Sergey Edunov, Michael Auli

With hundreds of languages used by people on our platforms and thousands more spoken around the world, developing powerful and flexible machine translation systems has long been a research focus for Facebook. Today we are proud to announce that Facebook AI models achieved first place in several language tasks included in this year's annual news translation competition, hosted by the Fourth Conference on Machine Translation (also known as WMT). Our models outperformed all other entrants' models in the four tasks we participated in, including English to German, the most competitive task in the contest, with entries drawn from a wide range of high-performing research teams. For this language direction, our translations have been declared superhuman by the WMT organizers, meaning that human evaluators preferred them over translations done by human experts.

- Yet...

Neural Machine Translation

You recently notified us of the possibility that copyrighted material was being made **available** through our website.

Sie haben uns vor Kurzem von der Überzeugung in Kenntnis gesetzt, dass urheberrechtlich geschütztes Material auf unserer Website **kostenlos verfügbar** ist.

You recently notified us of a belief that copyrighted material was being made **available at no cost** through our website.

https://unbabel.com/blog/machine-translation-customer-service/

Neural Machine Translation

If you live just **20 kilometres** away from San Diego, you may consider driving to the Westfield Mission Valley mall and collecting it yourself.

Si vous habitez à seulement **20 milles** de San Diego, vous pouvez envisager de vous rendre au centre commercial Westfield Mission Valley et de le récupérer vous-même.

If you live just **20 miles** from San Diego, you may consider driving to the Westfield Mission Valley mall and collecting it yourself.

https://unbabel.com/blog/machine-translation-customer-service/

Neural Machine Translation

It looks like it took a while for the subscription to be marked inactive **but it is cancelled now**.

Es scheint, dass es eine Weile gedauert hat, bis das Abonnement als inaktiv markiert wurde.

It looks like it took a while for the subscription to be marked inactive.

Neural Machine Translation

It looks like it took a while for the subscription to be marked inactive **but it is cancelled now**.

Es scheint, dass es eine Weile gedauert hat, bis das Abonnement als inaktiv markiert wurde.

It looks like it took a while for the subscription to be marked inactive.

The contract is understandable.

Le contrat est compréhensible, **veuillez nous appeler dès que possible**.

The contract is understandable, **please call us as soon as possible**.

https://unbabel.com/blog/machine-translation-customer-service/

Neural Machine Translation

Packages 1 and 2 both charge a monthly fee, as these have additional features to **Package** 1.

Pakketten 1 en 2 vragen elk een maandelijks bedrag, omdat deze extra functies hebben voor **Pakket** 1.

Abonnements 1 and 2 both charge a monthly fee, as these have additional features to **Abonnement** 1.

https://unbabel.com/blog/machine-translation-customer-service/

Ways to improve on NMT

In all cases: **very fluent MT!**

Ways to improve on NMT

In all cases: **very fluent MT!**

- Better NMT!
  - Document-wide translation
  - Coverage mechanism
  - External knowledge (e.g. terminology)
  - ...

Ways to improve on NMT

In all cases: **very fluent MT!**

- Better NMT!
  - Document-wide translation
  - Coverage mechanism
  - External knowledge (e.g. terminology)
  - ...
- Predicting quality to inform users: **Quality estimation**
- Fixing the NMT output: **Automatic post-editing**

Outline

Quality Estimation

- **QE**: metrics that provide an **estimate** on the **quality** of translations *on the fly*

Quality Estimation

- **QE**: metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality Estimation

- **QE**: metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

Quality Estimation

- **QE**: metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**

Quality Estimation

- **QE**: metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**

Quality = **How much effort to fix it?**

Quality Estimation

- **QE**: metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

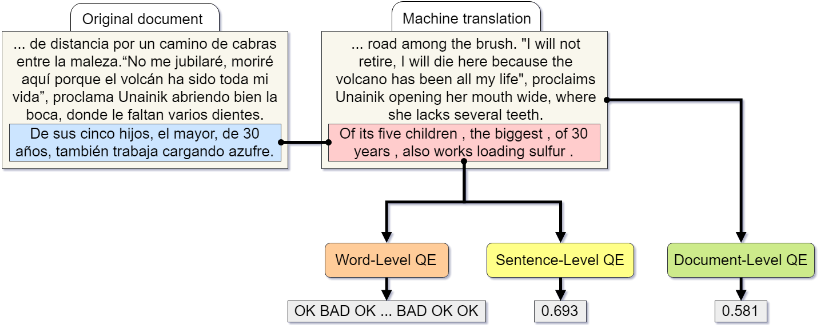
Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**

Quality = **How much effort to fix it?**

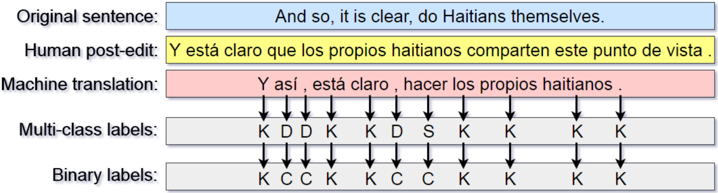
Quality = **Which words need fixing?**

Levels of granularity



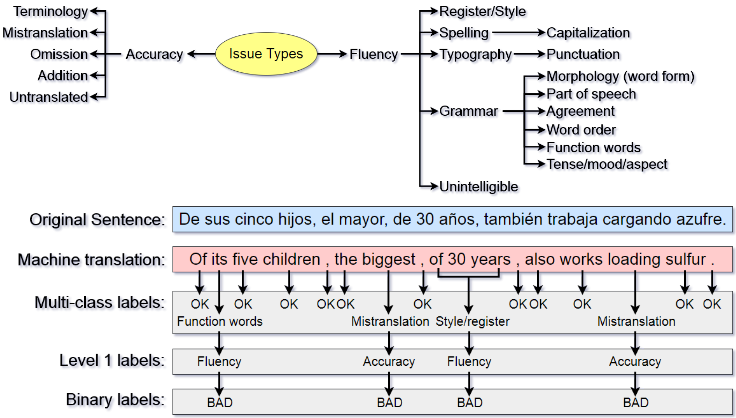
Word-level QE: labels

- Predict binary **GOOD/BAD** labels
- Predict general **types of edits**: replace, delete, keep

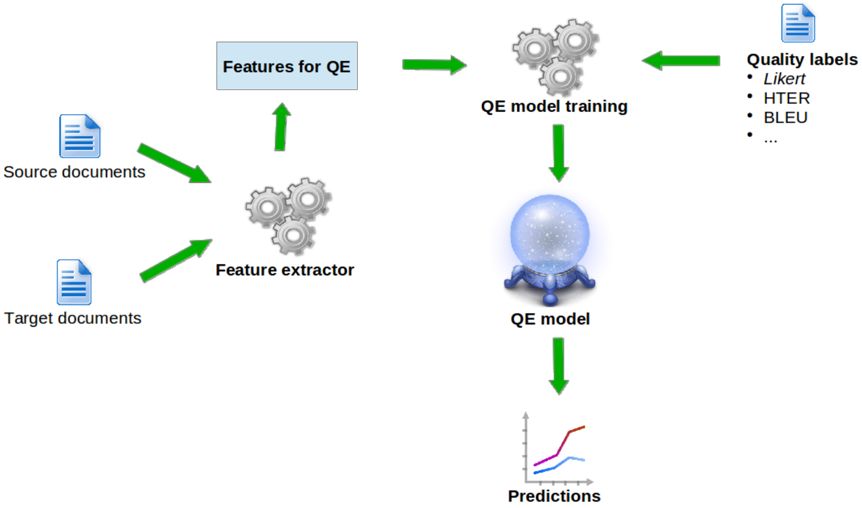


Word-level QE: labels

- Predict specific errors. E.g. **MQM**



“Traditional” framework



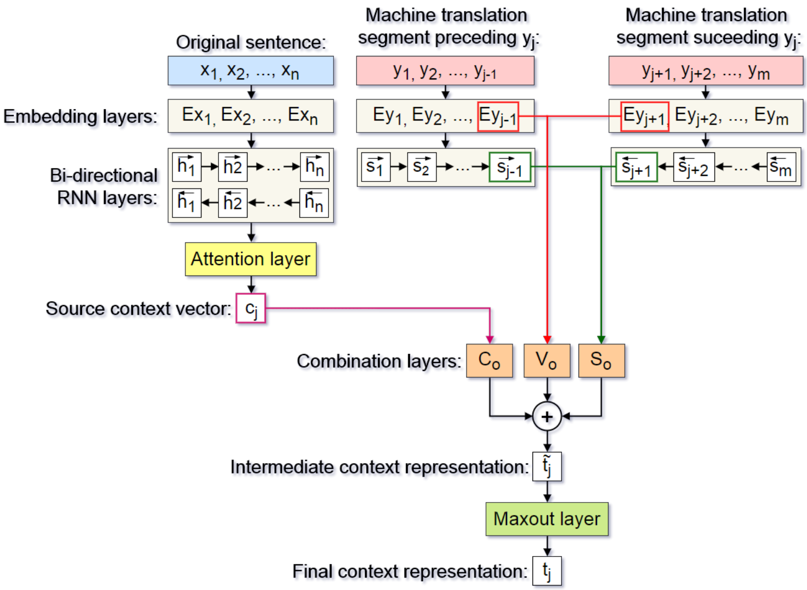
"Neural" framework

POSTECH model

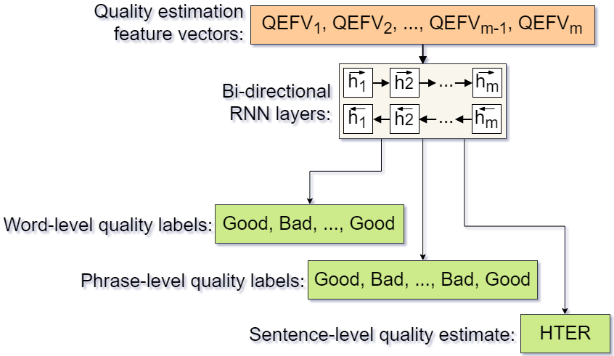
- **Predictor-estimator** sequential approach
  - **Predictor**: encoder-decoder RNN model to predict words based on their context, generating representations of good translations
  - **Estimator**: RNN model to produce quality estimates for words, phrases and sentences

Hyun Kim, Jong-Hyeok Lee and Seung-Hoon Na. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. WMT17

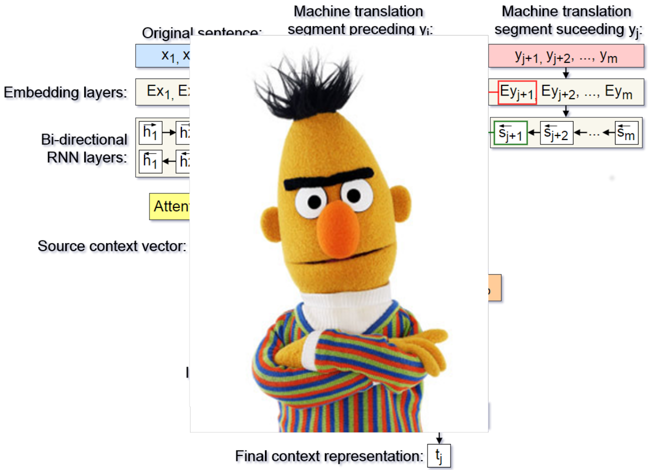
“Neural” framework - Predictor



"Neural" framework - Estimator



"Neural" framework - BERT/XLM/Laser predictor



“Neural” framework with BERT/XLM/Laser

Unbabel’s predictor-estimator

- Transformer-based predictor
- Transfer learning: pre-trained language models as feature extractors: multilingual BERT and XLM
- Fine-tuning: continuing predictor LM training on in-domain data

QE - sentence-level SOTA (WMT18)

Predicting HTER, **English–German** SMT:

Model	**Pearson r**
SMT DATASET	
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.74
QEBrain DoubleBi w/ BPE-tok	0.73
UNQE	0.70
TSKQE2	0.49
SHEF-PT	0.49
TSKQE1	0.48
UTartu/QuEst+Attention	0.43
UTartu/QuEst+Att+CrEmb3	0.42
sMQE	0.40
RTM_MIX7	0.39
RTM_MIX6	0.39
SHEF-bRNN	0.37
BASELINE	0.37

QE - sentence-level SOTA (WMT18)

Predicting HTER, **English–German** NMT:

	NMT DATASET
• UNQE	0.51
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.50
• QEBrain DoubleBi w/ word-tok	0.50
TSKQE1	0.42
TSKQE2	0.41
SHEF-bRNN	0.38
SHEF-PT	0.38
UTartu/QuEst+Attention	0.37
sMQE	0.37
UTartu/QuEst+Att+CrEmb3	0.37
BASELINE	0.29

QE - sentence-level SOTA (WMT19)

Predicting HTER, **English–German** NMT:

Model	Pearson
† UNBABEL Ensemble	0.5718
CMULTIMLT	0.5474
NJUNLP BiQE BERT Ensemble	0.5433
NJUNLP BiQE	0.5412
ETRI	0.526
Baseline	0.4001
UTARTU LABE	-0.319
UTARTU LABEL	0.2487
USAAR-DFKI CNNQE	0.2013
BOUN RTM1*	0.4734
BOUN RTM2*	0.1799

Same test set as 2018

QE - word-level SOTA (WMT18)

Predicting good/bad labels, **English-German** SMT vs NMT:

SMT DATASET Model	Words in MT F_1-BAD	 F_1-OK	 **F_1-mult**
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.68	0.92	0.62
QEBrain DoubleBi w/ word-tok	0.66	0.92	0.61
SHEF-PT	0.51	0.85	0.43
CMU-LTI	0.48	0.82	0.39
SHEF-bRNN	0.45	0.81	0.37
BASELINE	0.41	0.88	0.36
Doc2Vec	0.29	0.75	0.22
BagOfWords	0.28	0.73	0.20

NMT DATASET Model	Words in MT F_1-BAD	 F_1-OK	 **F_1-mult**
• QEBrain DoubleBi w/ word-tok (using voting)	0.48	0.91	0.44
• QEBrain DoubleBi w/ word-tok	0.48	0.92	0.43
CMU-LTI	0.36	0.85	0.30
SHEF-bRNN	0.35	0.86	0.30
SHEF-PT	0.34	0.87	0.29
BASELINE	0.20	0.92	0.18

QE - word-level SOTA (WMT19)

Predicting good/bad labels, **English-German** NMT:

Model	$\mathbf{F_1}$
† UNBABEL Ensemble	0.4752
UNBABEL Stacked	0.4621
ETRI BERT Multitask A	0.4061
ETRI BERT Multitask B	0.4047
MIPT Neural CRF Transformer	0.3285
MIPT Neural CRF RNN	0.3025
Baseline	0.2974
BOUN RTM GLMd*	0.1846

Same data as 2018, F1 = F1-mult

What helps - winning submission (WMT19)

Ensembling of multiple models, **English-German** NMT:

SYSTEM	TARGET F_1	SOURCE F_1	PEARSON
LINEAR	0.3346	0.2975	-
APE-QE	0.3740	0.3446	0.3558
APE-BERT	0.4244	0.4109	0.3816
PREDEST-RNN	0.3786	-	0.5020
PREDEST-TRANS	0.3980	-	0.5300
PREDEST-XLM	0.4144	0.3960	0.5810
PREDEST-BERT	0.3870	0.3310	0.5190
LINEAR ENS.	0.4520	0.4116	-
(*)POWELL'S ENS.	0.4872	0.4607	0.5968

Conclusions - QE

- Predicting NMT quality is **harder than for SMT**
  - Higher quality of NMT
  - Lower predictability of errors
- Performance on **general data** not clear
  - Task to predict DArr scores on news data at WMT19

	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	85,365	38,307	31,139	27,094	21,862	46,172	31,070
BEER	0.128	0.283	0.260	0.421	0.315	0.189	0.371
YISI-1_SRL	**0.199**	**0.346**	**0.306**	**0.442**	**0.380**	**0.222**	**0.431**
QE as a Metric:							
IBM1-MORPHEME	−0.074	0.009	–	–	0.069	–	–
IBM1-POS4GRAM	−0.153	–	–	–	–	–	–
LASIM	−0.024	–	–	–	–	0.022	–
LP	−0.096	–	–	–	–	−0.035	–
UNI	0.022	0.202	–	–	–	0.084	–
UNI+	0.015	0.211	–	–	–	0.089	–
YISI-2	0.068	0.126	−0.001	0.096	0.075	0.053	0.253
YISI-2_SRL	0.068	–	–	–	–	–	0.246
newstest2019							

Conclusions - QE

- Predicting NMT quality is **harder than for SMT**
  - Higher quality of NMT
  - Lower predictability of errors
- Performance on **general data** not clear
  - Task to predict DArr scores on news data at WMT19

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	85,365	38,307	31,139	27,094	21,862	46,172	31,070
BEER	0.128	0.283	0.260	0.421	0.315	0.189	0.371
YISI-1_SRL	**0.199**	**0.346**	**0.306**	**0.442**	**0.380**	**0.222**	**0.431**
QE as a Metric:							
IBM1-MORPHEME	−0.074	0.009	–	–	0.069	–	–
IBM1-POS4GRAM	−0.153	–	–	–	–	–	–
LASIM	−0.024	–	–	–	–	0.022	–
LP	−0.096	–	–	–	–	−0.035	–
UNI	0.022	0.202	–	–	–	0.084	–
UNI+	0.015	0.211	–	–	–	0.089	–
YISI-2	0.068	0.126	−0.001	0.096	0.075	0.053	0.253
YISI-2_SRL	0.068	–	–	–	–	–	0.246

newstest2019

- **Word-level prediction** really important, even harder!

Conclusions - QE

- Predicting NMT quality is **harder than for SMT**
  - Higher quality of NMT
  - Lower predictability of errors
- Performance on **general data** not clear
  - Task to predict DArr scores on news data at WMT19

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	85,365	38,307	31,139	27,094	21,862	46,172	31,070
BEER	0.128	0.283	0.260	0.421	0.315	0.189	0.371
YISI-1_SRL	**0.199**	**0.346**	**0.306**	**0.442**	**0.380**	**0.222**	**0.431**
QE as a Metric:							
IBM1-MORPHEME	−0.074	0.009	–	–	0.069	–	–
IBM1-POS4GRAM	−0.153	–	–	–	–	–	–
LASIM	−0.024	–	–	–	–	0.022	–
LP	−0.096	–	–	–	–	−0.035	–
UNI	0.022	0.202	–	–	–	0.084	–
UNI+	0.015	0.211	–	–	–	0.089	–
YISI-2	0.068	0.126	−0.001	0.096	0.075	0.053	0.253
YISI-2_SRL	0.068	–	–	–	–	–	0.246

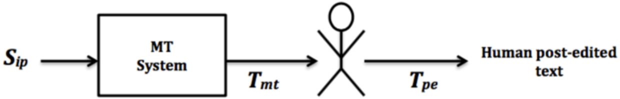
newstest2019

- **Word-level prediction** really important, even harder!
- **Usefulness** on any level still to be investigated

Outline

Task

Automatically fix the MT output:

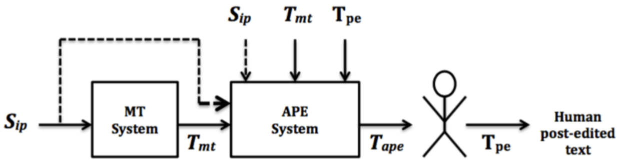


- **Goals**:
  - Minimise human post-editing effort
  - Assume MT system is black-box
  - Adapt general MT system to domain or translator

Figures by Santanu Pal

Task

Automatically fix the MT output:



- **Goals**:
  - Minimise human post-editing effort
  - Assume MT system is black-box
  - Adapt general MT system to domain or translator

Figures by Santanu Pal

Approach

Translate from *broken* into good target language

- Trained using post-edited data: <source, MT, PE> or <MT, PE>
- Same neural models as MT, but:
  - **Source** taken into account - multi-source models
  - Pre-trained representations for encoder/decoder (BERT)
  - Data augmentation with synthetic errors, often via **back-translation**

Approach

Back-translation

SRC	In patients with chronic renal failure, there is a pre-disposing development of metabolic acidosis.
MT_{EN-DE}	Bei Patienten mit chronischer Niereninsuffizienz kommt es vorab zu einer metabolischen Azidose.
MT_{DE-EN}	In patients with chronic renal insufficiency, metabolic acidosis occurs in advance.

Approach

Other strategies for **data augmentation**:

- Randomly generate HTER operations to match stats of original APE data
- Word-level QE to generate substitutions and deletions
- Replicate errors from the APE data if 1-2 right/left word contexts are the same
- Replicate trivial errors like missing quotes (grammar correction) - **works best**

Baseline and evaluation

Baseline:

- "Do nothing" - keep MT as is

Evaluation:

- **TER** $\downarrow$: edit distance between APE-fixed MT and human PE

Performance

- **1st generation**: monolingual SMT models to fix RBMT $\rightarrow$ very effective
- **2nd generation**: monolingual NMT/SMT models to fix SMT $\rightarrow$ effective enough
- **Currently**: monolingual NMT models to fix NMT $\rightarrow$ not so effective....

Performance

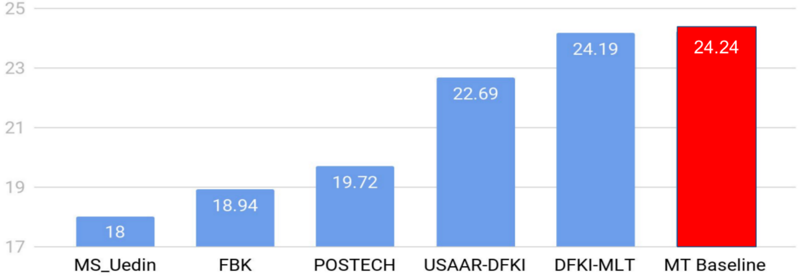
- **1st generation**: monolingual SMT models to fix RBMT $\rightarrow$ very effective
- **2nd generation**: monolingual NMT/SMT models to fix SMT $\rightarrow$ effective enough
- **Currently**: monolingual NMT models to fix NMT $\rightarrow$ not so effective....

Performance

- **1st generation**: monolingual SMT models to fix RBMT $\rightarrow$ very effective
- **2nd generation**: monolingual NMT/SMT models to fix SMT $\rightarrow$ effective enough
- **Currently**: monolingual NMT models to fix NMT $\rightarrow$ not so effective....

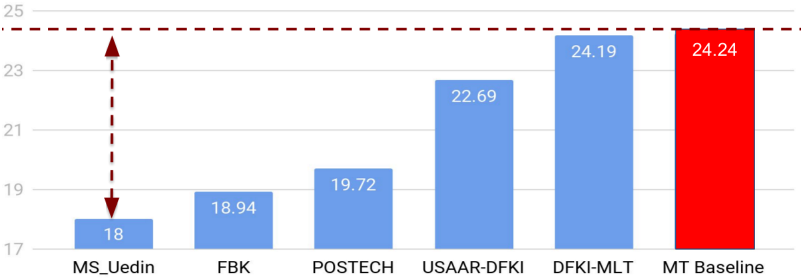
APE - SOTA (WMT18)

HTER for **English–German**, SMT:



APE - SOTA (WMT18)

HTER for **English–German**, SMT:

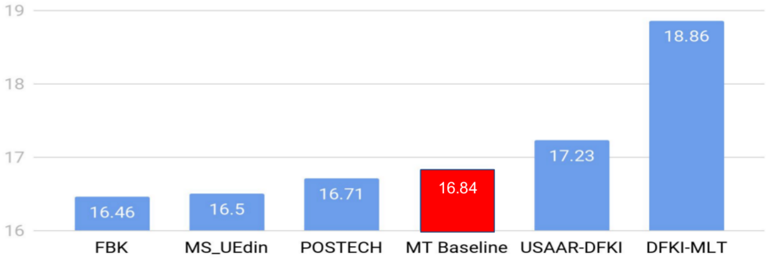


- Best system: -6.24 HTER (24.24 $\rightarrow$ 18.0)
- Steady increase: +0.3 in 2015, -3.24 in 2016, -4.88 in 2017, **-6.24 in 2018**

Figures from Findings of the Automatic Post-Editing Task at WMT18/19

APE - SOTA (WMT18)

HTER for **English–German**, NMT:

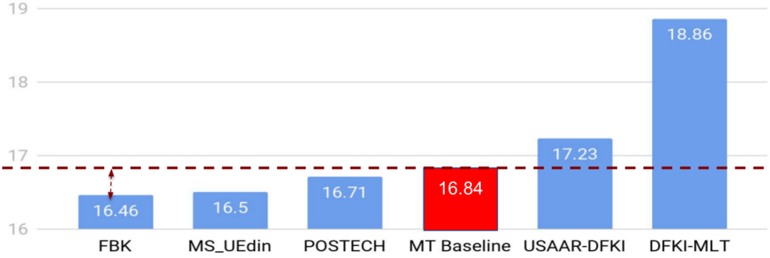


- Best system: -0.38 HTER (16.84 → 16.46)
- Correcting NMT output is a harder task

Figures from Findings of the Automatic Post-Editing Task at WMT18/19

APE - SOTA (WMT18)

HTER for **English–German**, NMT:

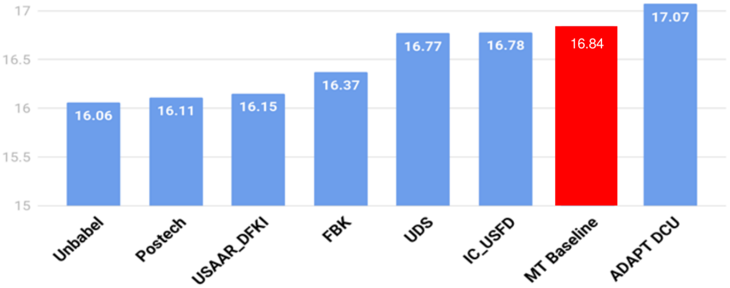


- Best system: -0.38 HTER (16.84 $\rightarrow$ 16.46)
- Correcting NMT output is a harder task

Figures from Findings of the Automatic Post-Editing Task at WMT18/19

APE - SOTA (WMT19)

HTER for **English–German**, NMT:

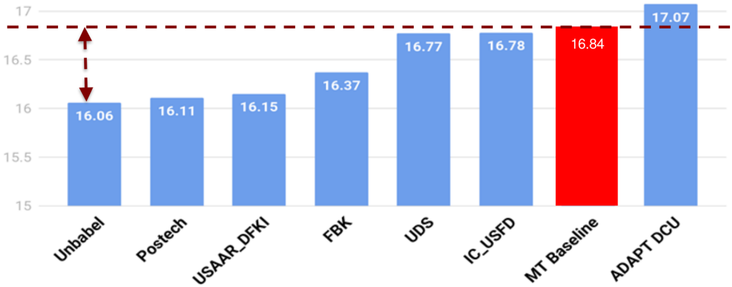


- Best system: -0.78 HTER (16.84 $\rightarrow$ 16.06)
- Same data... Correcting NMT output is still hard

Figures from Findings of the Automatic Post-Editing Task at WMT18/19

APE - SOTA (WMT19)

HTER for **English–German**, NMT:



- Best system: -0.78 HTER (16.84 $\rightarrow$ 16.06)
- Same data... Correcting NMT output is still hard

Figures from Findings of the Automatic Post-Editing Task at WMT18/19

APE - SOTA (WMT18-19)

Do systems **over or under-correct**?

- SMT: top systems modify **79-82%** of sentences
  - Expected: 85% of sentences
- NMT: top systems modify **4-39%** of sentences
  - Expected: 75% of sentences

Do systems make **right corrections**?

- SMT18: **55%** of corrections are right
- NMT18: **34%** of corrections are right
- NMT19: **45%** of corrections are right

Findings of the Automatic Post-Editing Task at WMT18/19

What helps in the winning submission

Unbabel's submission

- Encoder and decoder initialised with the pre-trained weights from **multilingual BERT**
- To avoid over-correction, **penalty** during beam decoding to constrain output to be as close as possible to input
- MT-REF pairs produced by in-domain training data to **augment APE data**

Conclusions - APE

- Fixing NMT is **harder than fixing SMT**
  - NMT quality is already very good
  - TER distribution in very skewed
  - NMT errors are less systematic, often “human-like”
- More promising for black-box, out-of-domain MT
- **Larger PE** datasets could help

Conclusions

- Machine translation is better but far from perfect
- **Quality estimation**
    - Can learn different types of quality
    - Important with NMT, but harder: NMT too fluent

Conclusions

- Machine translation is better but far from perfect
- **Quality estimation**
  - Can learn different types of quality
  - Important with NMT, but harder: NMT too fluent
- **Automatic post-editing** can still be useful
  - Systems under-correct and make fewer correct changes
  - Can be due to similarities in NMT and APE architectures

Conclusions

- Machine translation is better but far from perfect
- **Quality estimation**
  - Can learn different types of quality
  - Important with NMT, but harder: NMT too fluent
- **Automatic post-editing** can still be useful
  - Systems under-correct and make fewer correct changes
  - Can be due to similarities in NMT and APE architectures

- General or "out-of-domain" cases are more promising
- Upper and lowerbounds are better defined for APE
- **Utility of QE and APE** in practice - open questions

Quality Estimation and Automatic Post-editing in the Neural Machine Translation Era

Lucia Specia

Imperial College/University of Sheffield
l.specia@sheffield.ac.uk

HAT Workshop, Dublin, August 19th 2019