technische universität
dortmund

Unsupervised Quality Estimation for Neural Machine Translation

Frederik Polachowski

Supervisor: Bin Li

Machine Translation

Sources:
https://static.deepl.com/img/favicon/deepl_logo_600_300.png
https://de.wikipedia.org/wiki/Datei:Google_Translate_logo.svg
https://aws.amazon.com

technische universität
dortmund

Machine Translation / 2

Original	Jackson pidas seal kõne, öeldes, et James Brown on tema suurim inspiratsioon.
	Jackson gave a speech there saying that James Brown is his greatest inspiration.
Translation	Jackson gave a speech there, saying that his greatest inspiration is James Brown.
	Jackson made a speech there, saying that James Brown was his biggest inspiration.

Source: https://github.com/facebookresearch/mlqe

Table of contents

- Basics
- Dataset
- Methodology
- Results
- Discussion

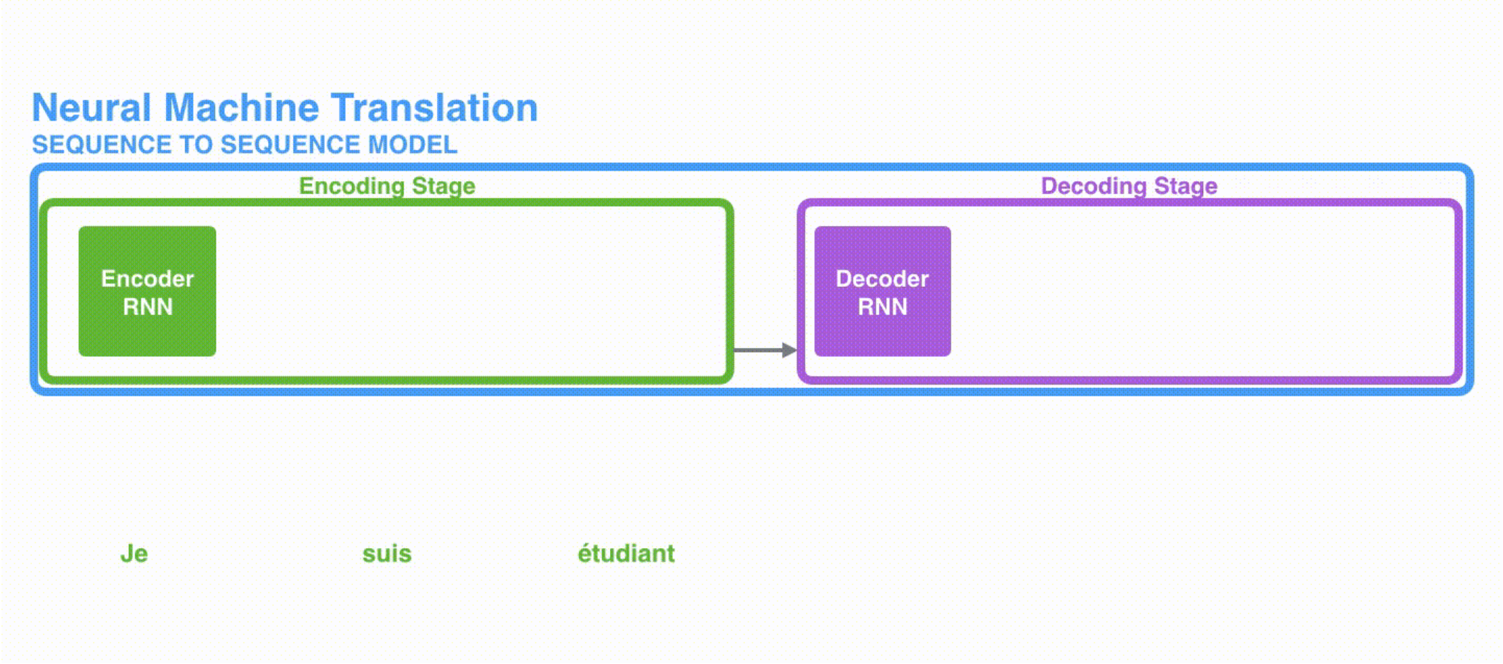
Basics – Meteor Similarity

- Evaluation of translation hypotheses with reference translations
- Calculation of sentence-level similarity scores
- Depending on the space of possible word alignments
    - Exact match
    - Stemming
    - Synonym
    - Paraphrases

Source: Denkowski & Lavie, *Meteor Universal: Language Specific Translation Evaluation for Any Target Language*, 2014

Basics – Transformer

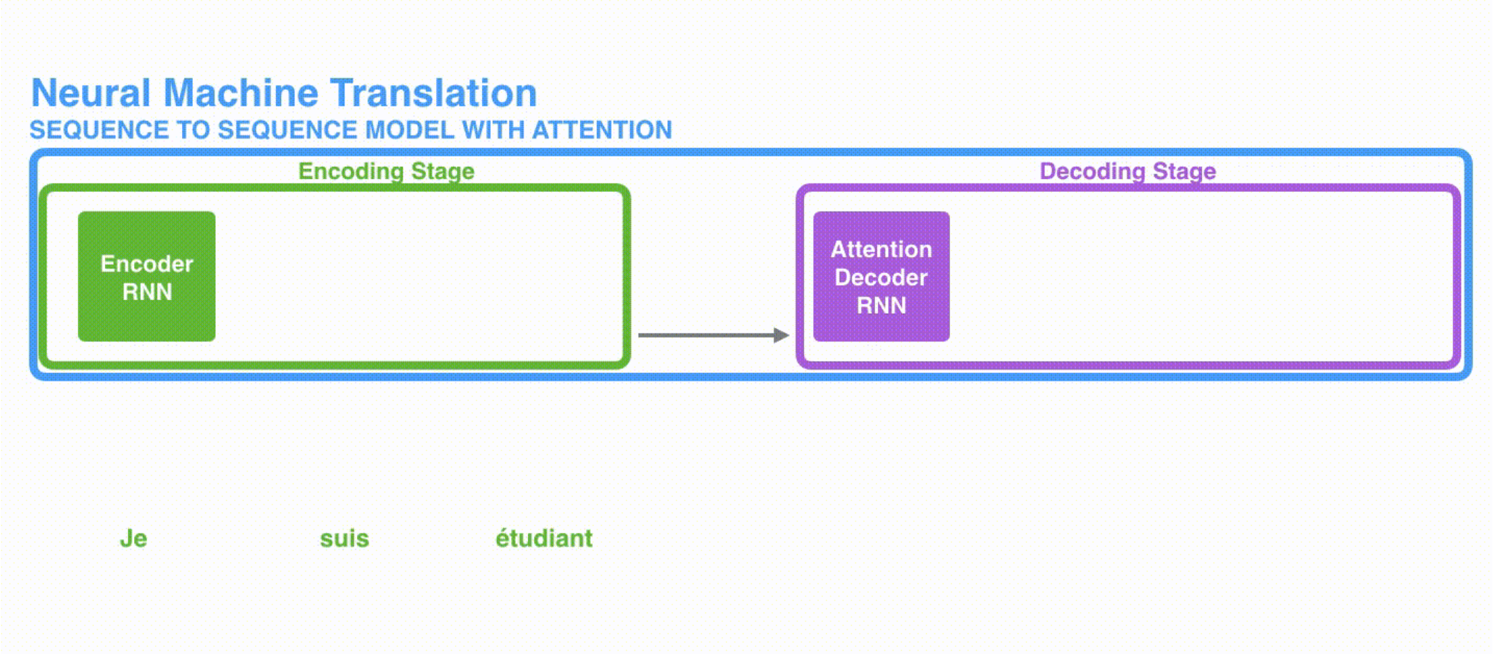
Seq-2-Seq Models



Source: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Basics – Transformer / 2

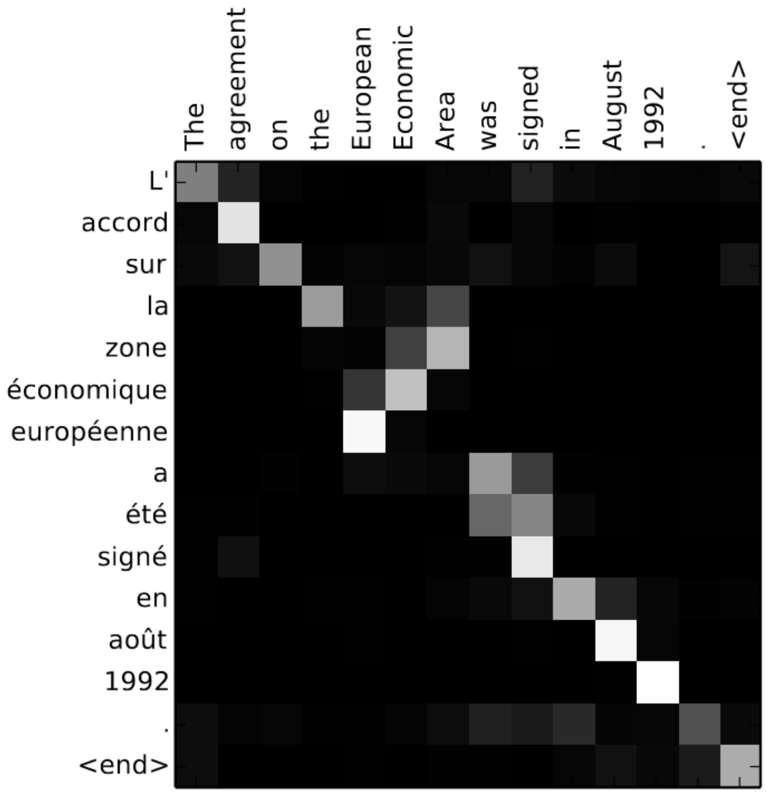
Attention



Source: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Basics – Transformer / 6

Attention Mechanism – word alignments



Source: Bahdanau et al., *Neural Machine Translation by Jointly Learning to Align and Translate*, 2015

Multilingual Dataset for QE

- 6 language pairs (EN, DE, ZH, Ro, Et, Si, Na)
- 10k sentences per language
- Scraped of Wikipedia in source languages
- Top 100 documents selected by
  - Intended source language
  - Between 50-100 character
  - Not contained in any other dataset
- ensured low-quality translation in test set

Multilingual Dataset for QE - Scoring

- Scoring based on direct assessment (DA)

- 6 annotators (from 2 different service provider)

- Each annotator rates the translation from 0 – 100
  - 0 – 10 incorrect
  - 11 – 29 few correct words
  - 30 – 50 major mistakes
  - 51 – 69 typos and grammatical errors but conveys meaning
  - 70 – 90 closely preserves semantics
  - 90 – 100 perfect translation

technische universität dortmund

Multilingual Dataset for QE – Scoring / 2

			scores				diff	
	pair	**size**	**avg**	**p25**	**median**	**p75**	**avg**	**std**
High-resource	**En-De**	23.7M	84.8	80.7	88.7	92.7	13.7	8.2
	En-Zh	22.6M	67.0	58.7	70.7	79.0	12.1	6.4
Mid-resource	**Ro-En**	3.9M	68.8	50.1	76.0	92.3	10.7	6.7
	Et-En	880k	64.4	40.5	72.0	89.3	13.8	9.4
Low-resource	**Si-En**	647k	51.4	26.0	51.3	77.7	13.4	8.7
	Ne-En	564k	37.7	23.3	33.7	49.0	11.5	5.9

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

Exploiting the Softmax Distribution

- Seq-to-Seq NMT architecture produces

$$p(y|x,\theta) = \prod_{t=1}^{T} p(y_t|y_{<t}, x, \theta)$$

- The sequence-level translation probability normalized by length

$$\mathrm{TP} = \frac{1}{T}\sum_{t=1}^{T} \log p(y_t|y_{<t}, x, \theta)$$

Exploiting the Softmax Distribution / 3

$$\mathrm{TP} = \frac{1}{T}\sum_{t=1}^{T} \log p(y_t | y_{<t}, x, \theta)$$

- the more confident the network the better the translation
- only 1-best probability estimates
- tend to be overconfident

Exploiting the Softmax Distribution / 4

$$\text{Softmax}{-}\text{Ent} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{v=1}^{V}\mathrm{p}(y_t^v)\log p(y_t^v) \qquad \text{with } p(y_t) = p(y_t|y_{<t}, x, \theta)$$

- sum over entire vocabulary V
- high quality if probability mass concentrated
- low quality if uniformly distributed
- but $[0.5, 0.5]$ and $[0.9, 0.1]$ produce the same mean

Exploiting the Softmax Distribution / 5

$$\text{Sent–Std} = \sqrt{\mathbb{E}[P^2] - \mathbb{E}[P]^2} \qquad \text{with } P = \log p(y_1), \ldots, \log p(y_T)$$

- calculate the standard-deviation of
- P, represents word-level log-probabilities

Exploiting the Softmax Distribution / 5

	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	0.399	0.482	<u>0.486</u>	<u>0.647</u>	0.208	0.257
Softmax-Ent	<u>0.457</u>	<u>0.528</u>	0.421	0.613	0.147	0.251
Sent-Std	0.418	0.472	0.471	0.595	<u>0.264</u>	<u>0.301</u>

Pearson(r) correlation between Softmax QE and human DA judgement.

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

technische universität dortmund

Quantifying Uncertainty

The goal is to approximate the posterior distribution that quantifies model uncertainty.

- perform N forward passes
- using Monte Carlo dropout

1. calculate mean and variance of posterior probabilities
2. compare the similarity of the output hypothesis

Quantifying Uncertainty – Posterior prob.

Mean

$$\mathrm{D{-}TP} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{TP}_{\theta' n} \qquad \text{with} \qquad TP = \frac{1}{T}\sum_{t=1}^{T} \log p(y_t | y_{<t}, x, \theta)$$

Variance

$$\mathrm{D{-}Var} = \mathbb{E}\left[\mathrm{TP}^2_{\theta'}\right] - \mathbb{E}[\mathrm{TP}_{\theta'}]^2$$

Combination of both

$$\mathrm{D{-}Combo} = (1 - \frac{\mathrm{D{-}TP}}{\mathrm{D{-}Var}})$$

Quantifying Uncertainty – Similarity score

$$\text{D-Lex-Sim} = \frac{1}{C}\sum_{i=1}^{|\mathbb{H}|}\sum_{j=1}^{|\mathbb{H}|} sim(h_i, h_j)$$

with $h_i, h_j \in \mathbb{H}, i \neq j$ and $C = \frac{|\mathbb{H}|(|\mathbb{H}|-1)}{2}$

and *Meteor* is used for similarity comparison

technische universität dortmund

Quantifying Uncertainty – Example

Low Quality	**Original**	Tanganjikast püütakse niiluse ahvenat ja kapentat.
	Reference	Nile perch and kapenta are fished from Lake Tanganyika.
	MT Output	There is a silver thread and candle from Tanzeri.
	Dropout	There will be a silver thread and a penny from Tanzer.
		There is an attempt at a silver greed and a carpenter from Tanzeri.
		There will be a silver bullet and a candle from Tanzer.
		The puzzle is being caught in the chicken's gavel and the coffin.
High quality	**Original**	Siis aga võib tekkida seesmise ja välise vaate vahele lõhe.
	Reference	This could however lead to a split between the inner and outer view.
	MT Output	Then there may be a split between internal and external viewpoints.
	Dropout	Then, however, there may be a split between internal and external viewpoints.
		Then, however, there may be a gap between internal and external viewpoints.
		Then there may be a split between internal and external viewpoints.
		Then there may be a split between internal and external viewpoints.

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

technische universität
dortmund

Quantifying Uncertainty

	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
D-TP	0.460	0.558	<u>0.642</u>	<u>0.693</u>	<u>0.259</u>	<u>0.321</u>
D-Var	0.307	0.299	0.356	0.332	0.164	0.232
D-Combo	0.286	0.418	0.475	0.383	0.189	0.225
D-Lex-Sim	<u>0.513</u>	<u>0.600</u>	0.612	0.669	0.172	<u>0.313</u>

Pearson(r) correlation between Uncertainty QE and human DA judgement.

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

Attention weights

“Attention weights represent the strength of connection between source and target token”

$$\text{Att−Ent} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_{ji}\log\alpha_{ji}$$

with α the attention weight
I the number of target tokens
J the number of source tokens

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

Attention weights / 2 – Multi head attention

$[H \times L]$ matrices of attention weights

$$\mathrm{AW{:}Ent{-}Min} = \min_{hl} \mathrm{Att{-}Ent}_{hl}$$

$$\mathrm{AW{:}Ent{-}Avg} = \frac{1}{H \times L} \sum_{h=1}^{H} \sum_{l=1}^{L} \mathrm{Att{-}Ent}_{hl}$$

technische universität dortmund

Attention weights / 3

	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
AW:Ent-Min	0.097	0.265	0.329	<u>0.524</u>	0.000	0.067
AW:Ent-Avg	0.1	0.205	0.377	0.382	0.090	0.112
AW:best head/layer	<u>0.255</u>	<u>0.381</u>	<u>0.416</u>	<u>0.636</u>	<u>0.241</u>	<u>0.168</u>

Pearson(r) correlation between Attention QE and human DA judgement.

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

Supervised QE

- PredEst Model
  - Encoder-decoder RNN – word predictor
  - Unidirectional RNN – quality estimator

- BiRNN Model
  - BERT Model – word predictor
  - Bidirectional RNN – source sentence encoder
  - Bidirectional RNN – target sentence encoder
  - Sigmoid layer – sentence-level quality estimator

Source:
Kim et. al., *Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation*, 2017
Blain et al. *Quality in, quality out: Learning from actual mistakes, 2020*

Supervised QE / 2

	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
BERT-BiRNN	**0.473**	**0.546**	**0.635**	**0.763**	**0.273**	**0.371**

Pearson(r) correlation between supervised QE and human DA judgement.

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

technische universität dortmund

Methodology – Comparison

	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
Sent-Std	0.418	0.472	0.471	0.595	**0.264**	0.301
D-Lex-Sim	**0.513**	**0.600**	0.612	0.669	0.172	**0.313**
AW:best head/layer	0.255	0.381	0.416	0.636	0.241	0.168
BERT-BiRNN	**0.473**	0.546	**0.635**	**0.763**	**0.273**	**0.371**

Source: Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation*, 2020

Future work

- Extend to other levels (word, phrase, document)
- Combined as features in supervised QE
- Different problem domain
  - Machine transcription
  - Semi-supervised labelling
  - Classification
  - Regression
- Quality measure for ensemble systems
- Information of translation quality in translation systems

Discussion – Not good aspects

- Sentences are rated in isolation
  - no context for information

- Non conform ratings are not truly rejected
  - they are repeated till “*consensus*”

- Rated by only two different sources of truth
  - done by *“professionals”*

Discussion – Good aspects

- High complexity of dataset
- Extensive result analysis
- Good visualization of important concepts / findings
- Validation of additional aspects

Thank you for your attention!

technische universität
dortmund

Bibliography

- Fomicheva et. al., *Unsupervised Quality Estimation for Neural Machine Translation.* Transactions of the Association for Computational Linguistics, 2020

- Bahdanau et al., *Neural Machine Translation by Jointly Learning to Align and Translate.* 3rd International Conference on Learning Representations, 2015

- Denkowski & Lavie, *Meteor Universal: Language Specific Translation Evaluation for Any Target Language.* Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014

- Kim et. al., *Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation.* Proceedings of the Second Conference on Machine Translation, pages 562–568, 2017

- Blain et al. *Quality in, quality out: Learning from actual mistakes.* Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2020

- http://jalammar.github.io