Neural Machine Translation: an Overview

Marco Dinarelli
Researcher of the National Council of Scientific Research (CNRS in France)
Laboratoire d'Informatique de Grenoble (LIG)
https://fr.wikipedia.org/wiki/Laboratoire_d%27informatique_de_Grenoble
Getalp group

Outline

- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- SMT/NMT Evaluation
- Document-Level NMT (CA-NMT)
- CA-NMT Evaluation
- Explainability
- Conclusions

A bit of symbols

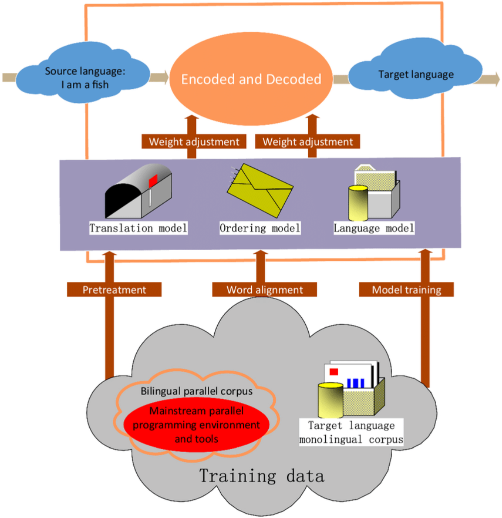
- x: input data, e.g. like in $y = f(x)$
- y: output data
- s: source symbol (input data as well)
- t: target symbol (output data)
- h: hidden state
- P: probability (model)

Same symbols in uppercase: sequences.
E.g. T: sequence of target symbols

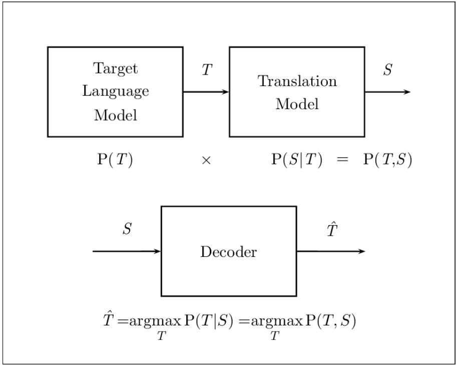
Same symbols with $\tilde{}$ or $\hat{}$: model's predictions (or *hypothesis*).
E.g. $\hat{T}$: model's translation for T.

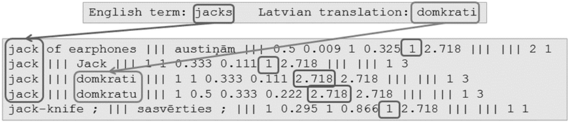
Statistical Machine Translation

The Dark Ages: Statistical Machine Translation (SMT)



SMT: more formally ...

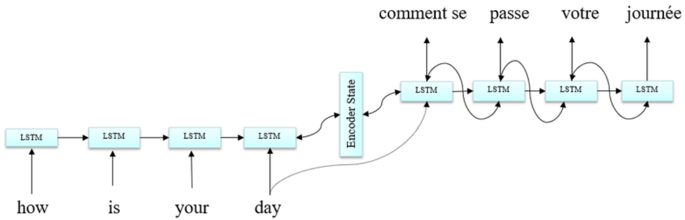


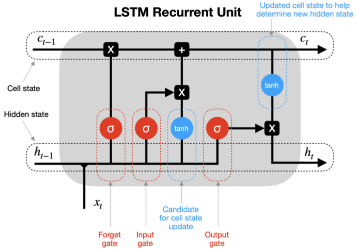


Source: https://www.researchgate.net

Neural Machine Translation

Neural Machine Translation (NMT): The Origin (2014)



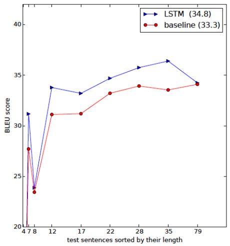


Source: https://www.researchgate.net

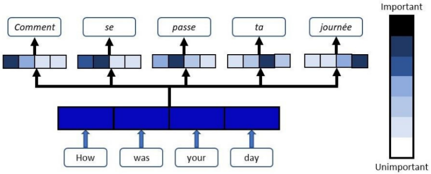
NMT (continued)

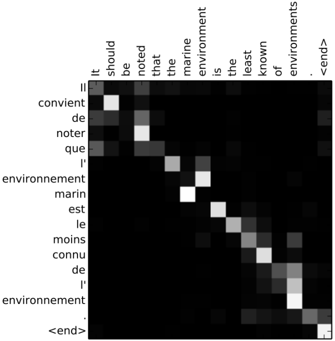
$$p(y_t | \{y_1, \ldots y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \tag{1}$$

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**



The Attention Mechanism (2014)





Source: https://teksands.ai

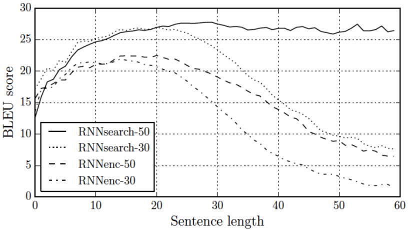
The Attention Mechanism (continued)

$$p(y_t | \{y_1, \ldots y_{t-1}\}, c_t) = g(y_{t-1}, s_t, c_t)$$

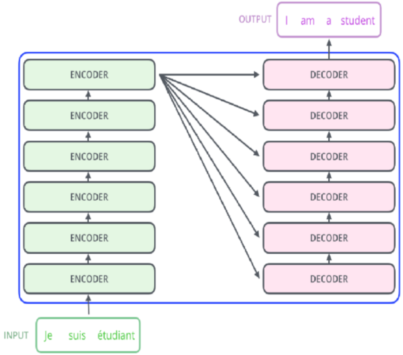
$$c_t = \sum_j \alpha_{i,j} h_j^e$$

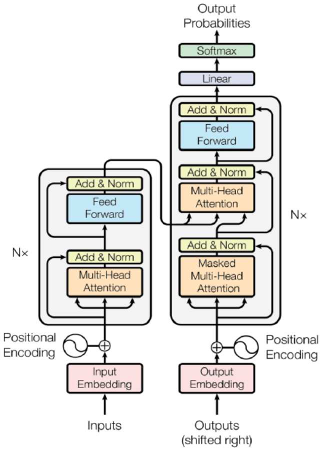
$$\alpha_{i,j} = \frac{\mathsf{exp}(a_{i,j})}{\sum_j \mathsf{exp}(a_{i,j})}$$

$$a_{i,j} = f(h_i^d, h_j^e)$$



The *Transformer* Model (2017)





The *Transformer* Model (continued)

The (self/cross) attention mechanism:

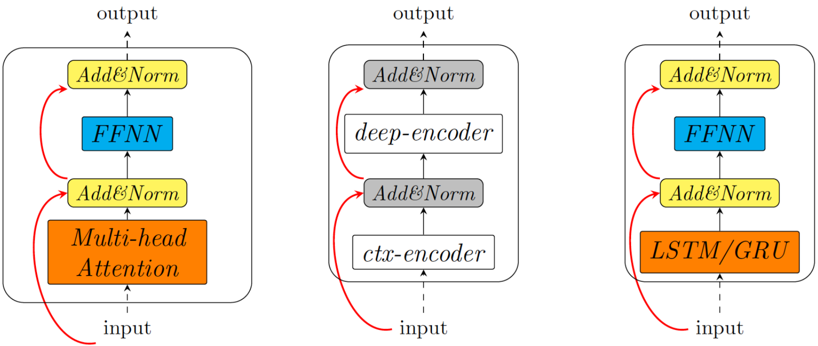
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The *Transformer* Model (continued)

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Encoder-Decoder Architecture (2014 - Present)



The contextual-encoder (ctx-encoder) can be any of:

- Recurrent layer (LSTM/GRU)
- Attention layer
- Convolutional layer

NMT: in summary

1. Conceptually (mathematically ?) simpler
   $P(y|h)$, that's all
   vs. SMT:
   $P(S,T) \times P(T); P(S,T) = \prod_{i=1}^{N} P(y_i|y_{i-1}, y_{i-2} \ldots x_1, \ldots x_M) \ldots$
2. Very effective:
   - SOTA in many domains
   - LLMs (AI!)
3. Less *explicit* behavior: "Black-box models"
   $\Rightarrow$ Explainability research axis
4. Examples of tools/systems:
   - SMT: Moses, Google translate
   - NMT: DeepL, Google translate (!)

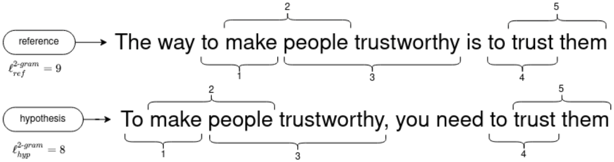
Evaluation

Evaluation Measure

- T: what you wanted the model to predict
  Reference (or *gold standard*, *ground truth*, whatever...)
- $\hat{T}$: what the model predicted
  Hypothesis
- Evaluation measure for MT: $f(T, \hat{T})$
  The higher the better (for most metrics...)

Evaluation Measure: n-gram matches

BLEU: **B**i-**L**ingual **E**valuation **U**nderstudy (2002)



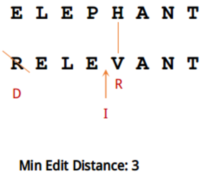
$$
\begin{aligned}
Geometric\ Average\ Precision\ (N) &= exp(\sum_{n=1}^{N} w_n\ log\ p_n) \\
&= \prod_{n=1}^{N} {p_n}^{w_n} \\
&= (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}
\end{aligned}
$$

Source: https://clementbm.github.io

Evaluation Measure: edit-distance based

TER: **T**ranslation **E**dit **R**ate (2006)
Same idea as edit distance (plus a shift)

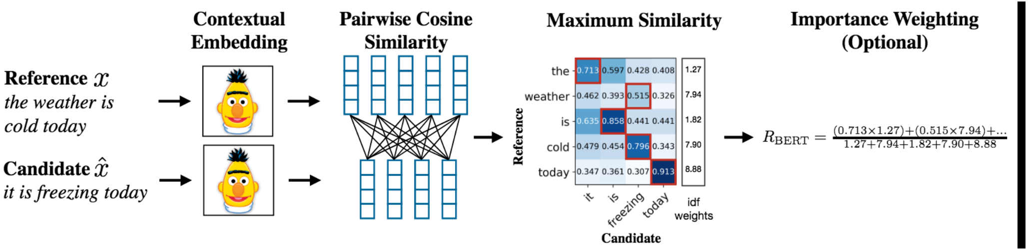
		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3



Source: https://www.ritambhara.in

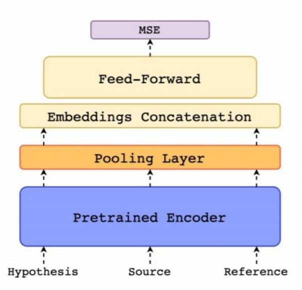
Evaluation Measure: deep embeddings

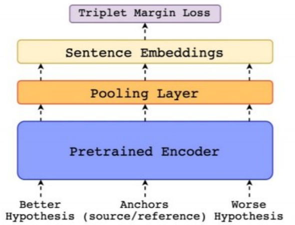
BERTscore (2020)
Similar idea as edit distance, but tokens are deep representations



Evaluation Measure: learned "metrics"

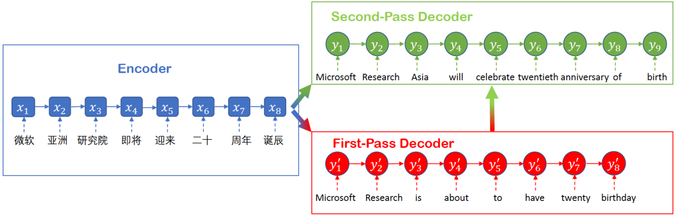
COMET: **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation (2020)
1. Predicts human judgments
2. Ranks "better" hypotheses
3. There's a version without *Reference* (QE)





The “*Human-Level Quality*” Debate (2018)

Hassan et al. (2018) paper: “Achieving Human Parity on Automatic Chinese to English News Translation”
→ raised the debate



Toral et al. (2018) paper: “Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation”
→ *critized* Hassan’s et al. evaluation method
→ basically gave birth to *Document-Level* NMT !

Document-Level Neural Machine Translation

Document-Level or Context-Aware NMT ?

- Document-Level means the whole document is used as context
- In practice: few sentences are used as context
  - Is the rest relevant ?
  - Computationally feasible
    $\rightarrow$ beyond LLMs

$\Rightarrow$ **Context-Aware NMT (CA-NMT)**

CA-NMT: two main (specific) solutions

- Concatenation models
- Multi-encoder models

\+

- LLMs (not specific)

CA-NMT: concatenation approach

- Standard *Transformer* architecture
- Just take N concatenated source/target sentences

+10

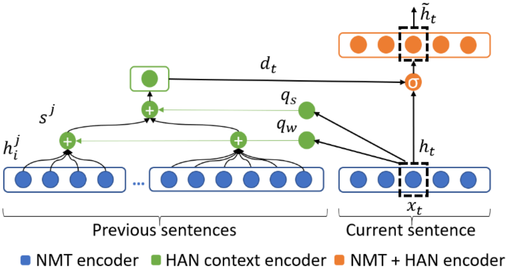
1 2 3 4 15 16 17 18 19

Hey there ! <S> How are you ? <E>

$$\mathrm{CD} \cdot \mathcal{L}_{context} + \mathcal{L}_{current}$$

CA-NMT: multi-encoder approach

- *Transformer* with additional attention mechanisms



CA-NMT Evaluation

Traditional quantitative evaluation

NMT model / Metric	*BLEU*	*COMET*	*ChrF*	*TER*
Multi-encoder	32.17	0.83	59.04	56.53
Concat*	32.08	0.81	58.62	57.38

Contrastive Test suites

source sentence with antecedent	*What's with the door?*
target sentence with antecedent	*Was ist mit der Tür?*
source context	***It*** *won't open.*
reference context	***Sie*** *geht nicht auf.*
source sentence	*- Is* ***it*** *locked?*
reference sentence	*- Ist* ***sie*** *abgeschlossen?*
contrastive 1	*- Ist* ***er*** *abgeschlossen?*
contrastive 2	*- Ist* ***es*** *abgeschlossen?*

Source: https://www.researchgate.net

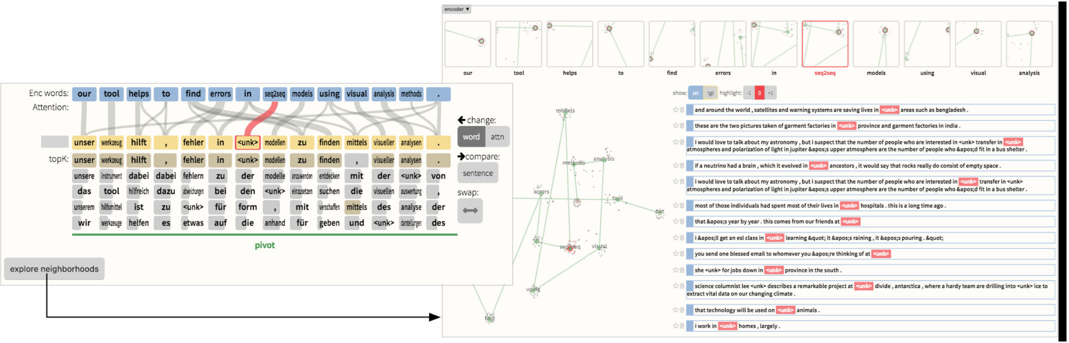
CA-NMT Evaluation with Contrastive Test Suites

NMT model	*ContraPro Accuracy*
Baseline	45.00
(Zhang et al., 2018)	42.60
*(Tu et al. 2018)**	45.20
(Muller et al., 2018a) concat21	48.00
(Muller et al., 2018a) concat22*	70.80
*(Maruf et al., 2019)**	39.15
(Voita et al., 2018)	42.55
(Stovanojski et al., 2019)	52.55
*(Muller et al., 2018b)** best	58.13
Multi-encoder	61.09
Concat*	74.39

Explainability

Neural models are (very) powerful...

But are they explainable ?
$\rightarrow$ *Black box* models



Main research axes

- **Faithfullness**: make the model's predictions coherent with its behavior

- **Plausibility**: are model's predictions explainable by its behavior ?

Our contribution within MAKENMT-Viz

“Context-Aware Neural Machine Translation Analysis and Evaluation Through Attention”. *Dinarelli et al.*, French journal TAL, 2024.

The idea: providing an explicit evaluation on discourse phenomena
How ? Using attention weights over coreference links

Our contribution (continued)

Data: ParCorFull 2.0
Parallel corpus (English, French, German, Portuguese) annotated with coreferences

The procedure:

- Translate the data with CA-NMT (En-De)
- Align CA-NMT input/output with corpus input/output
- Score coreference links (attention weights)

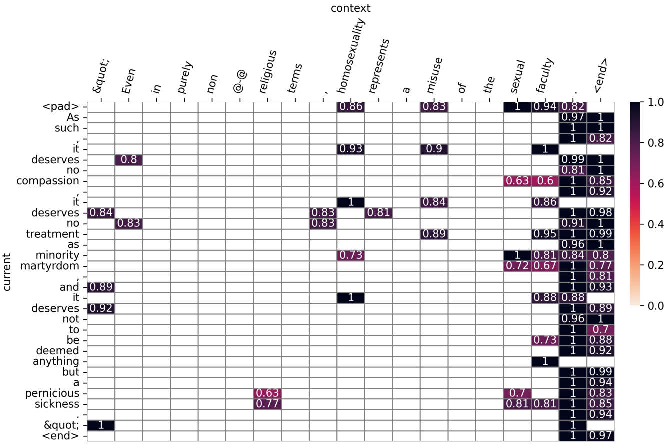
Our contribution (continued)

Two evaluation: quantitative.

NMT model / Metric	Max-weight	Non-zero weight	Average weight
Multi-encoder (src)	45.91%	88.83%	0.8183
Concat (src)	10.45%	50.98%	0.2994
Concat (tgt)	13.25%	33.22%	0.2136

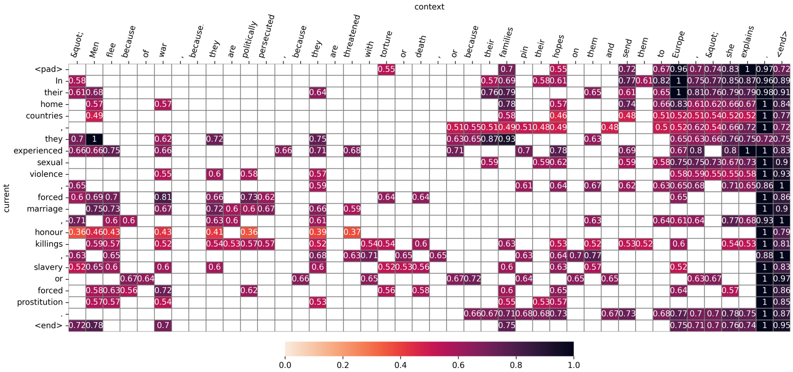
Our contribution (continued)

Two evaluation: qualitative.



Our contribution (continued)

Qualitative evaluation: an interesting example 1/2:



Our contribution (continued)

Qualitative evaluation: an interesting example 2/2: