# OpenNMT Evaluation : State of the Art ?

Vincent Nguyen

OpenNMT Workshop Paris, March 2018

# Who are we ?

- Founded in 1991 in Paris, now established in France, UK, Belgium Spain, USA, Canada
- Revenue : 70 Million Euros
  - 45% Translation
  - 30% Summarization
  - 20% Transcription
  - 5% not NLP
- Small but very active R&D team

[ubiqus]

# OpenNMT Evaluation

- Benchmarking toolkits is a very difficult task, which requires:
  - Extensive knowledge and efforts to reproduce other systems' results
  - Being thorough to set-up comparable contexts
  - Good faith / independence

- OpenNMT is very competitive in terms of results, speed, … and the biggest user community!

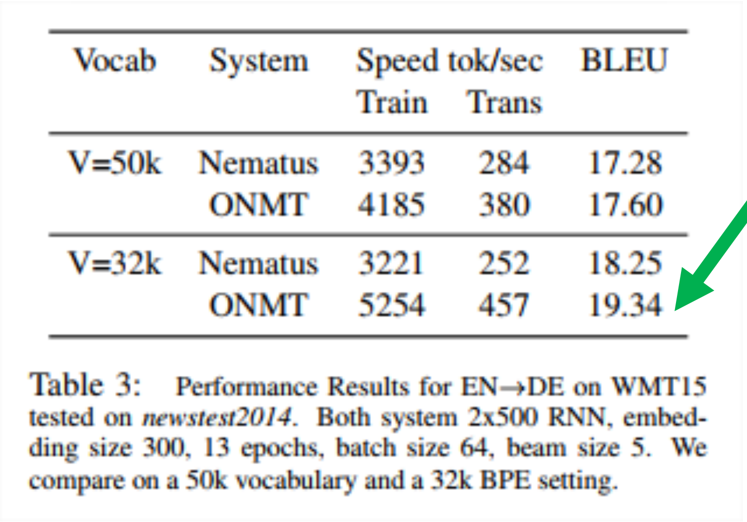
- OpenNMT will keep pace with innovative architecture, and facilitate usage (nmt-wizard)

# What is BLEU Evaluation ?

- 0-100 scale which measures how close the MT output is, compared to a human reference
  - < 40 : requires heavy post-editing
  - > 40 < 60 : requires light post-editing
  - > 60 : almost perfect very light proofreading
- However no measure is perfect
  - Even a BLEU score of 60 is about 25 TER (translation edit rate) on a per word basis, and 21 Levenshtein distance on per character basis
- But even a human reference is just one reference

# It all started with …. our own « poor » results

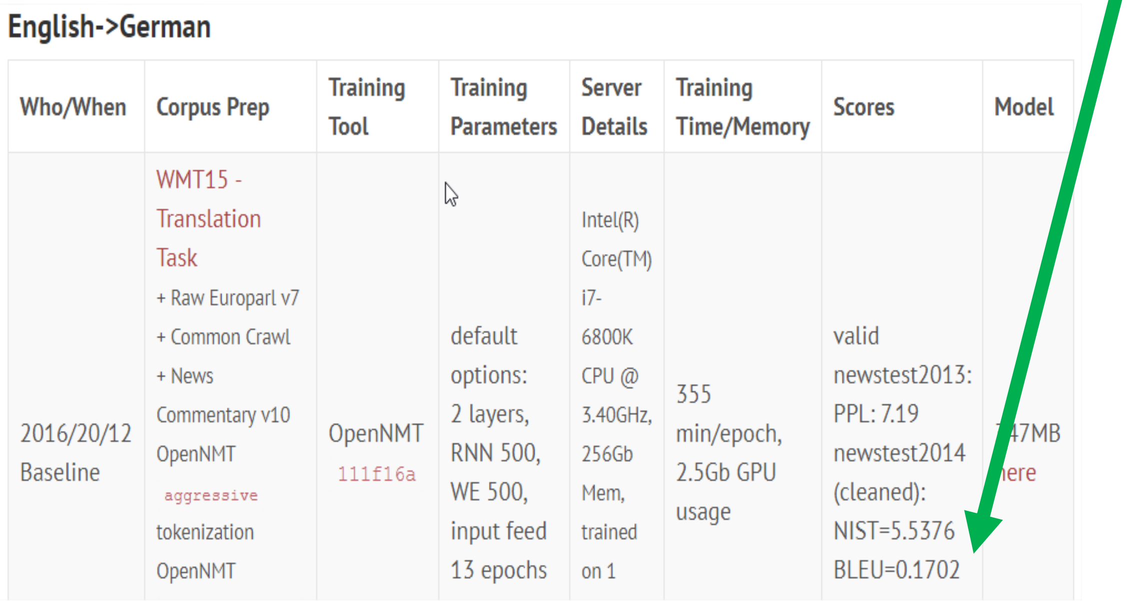
- OpenNMT: open-source Toolkit paper (march 2017)



| Vocab | System | Speed tok/sec Train | Speed tok/sec Trans | BLEU |
|---|---|---|---|---|
| V=50k | Nematus | 3393 | 284 | 17.28 |
| | ONMT | 4185 | 380 | 17.60 |
| V=32k | Nematus | 3221 | 252 | 18.25 |
| | ONMT | 5254 | 457 | 19.34 |

Table 3: Performance Results for EN→DE on WMT15 tested on *newstest2014*. Both system 2x500 RNN, embedding size 300, 13 epochs, batch size 64, beam size 5. We compare on a 50k vocabulary and a 32k BPE setting.

- ONMT Web site:



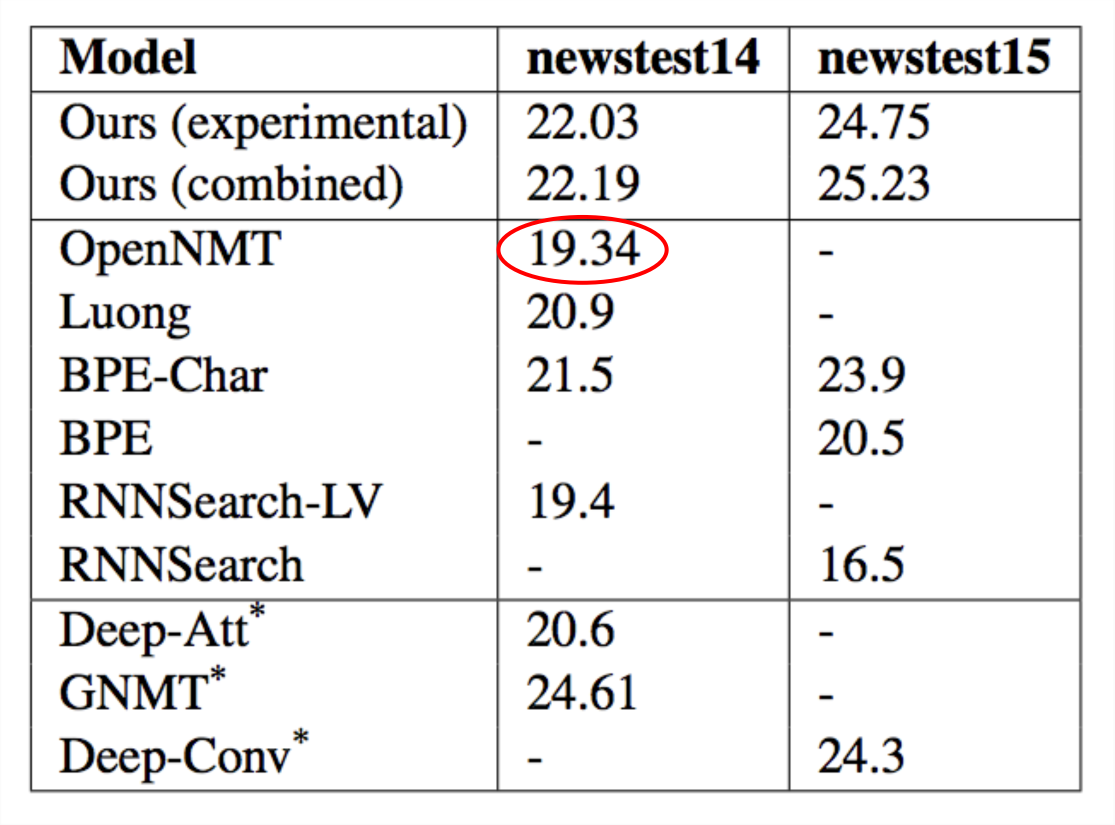
**English->German**

| Who/When | Corpus Prep | Training Tool | Training Parameters | Server Details | Training Time/Memory | Scores | Model |
|---|---|---|---|---|---|---|---|
| 2016/20/12 Baseline | WMT15 - Translation Task<br>+ Raw Europarl v7<br>+ Common Crawl<br>+ News Commentary v10<br>OpenNMT `aggressive` tokenization<br>OpenNMT | OpenNMT `111f16a` | default options:<br>2 layers, RNN 500, WE 500, input feed<br>13 epochs | Intel(R) Core(TM) i7-6800K CPU @ 3.40GHz, 256Gb Mem, trained on 1 | 355 min/epoch, 2.5Gb GPU usage | valid newstest2013: PPL: 7.19<br>newstest2014 (cleaned): NIST=5.5376 BLEU=0.1702 | 47MB<br>here |

- We published « baseline » systems and not « state-of-the-art » comparable results

# Then literature reported « results » …

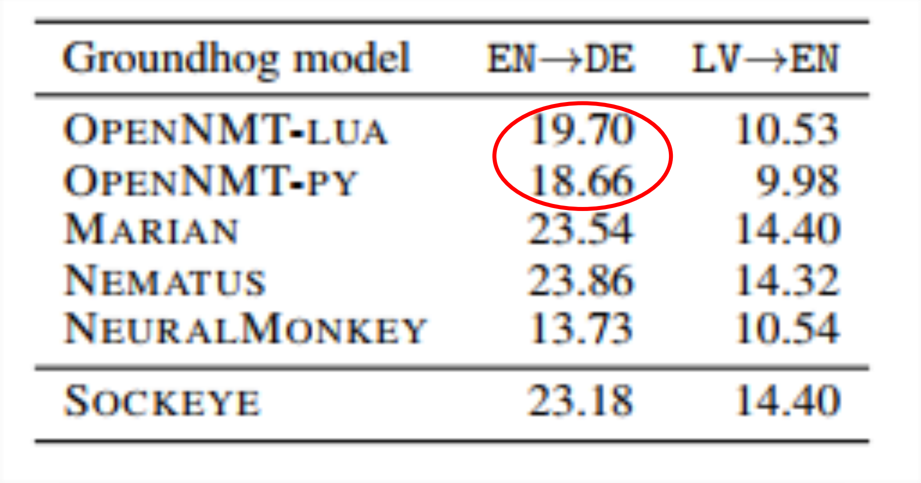
- Denny Britz: Massive exploration of NMT (2017)



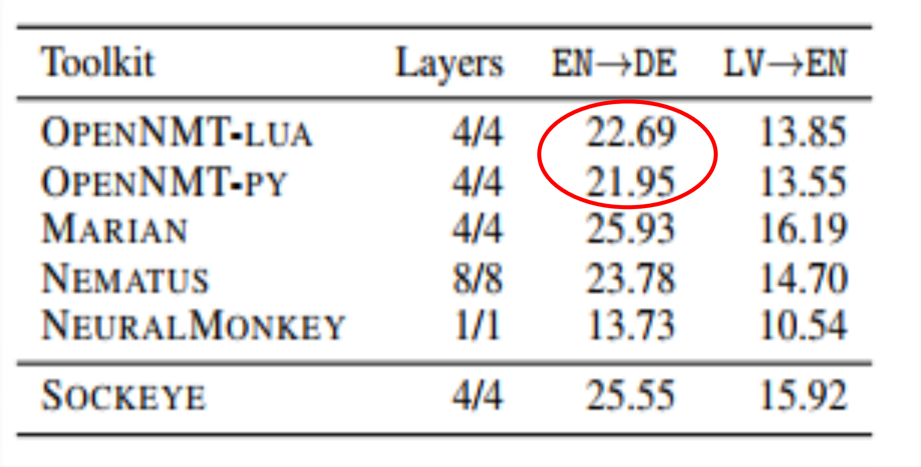
| Model | newstest14 | newstest15 |
|---|---|---|
| Ours (experimental) | 22.03 | 24.75 |
| Ours (combined) | 22.19 | 25.23 |
| OpenNMT | 19.34 | - |
| Luong | 20.9 | - |
| BPE-Char | 21.5 | 23.9 |
| BPE | - | 20.5 |
| RNNSearch-LV | 19.4 | - |
| RNNSearch | - | 16.5 |
| Deep-Att* | 20.6 | - |
| GNMT* | 24.61 | - |
| Deep-Conv* | - | 24.3 |

- Not very enticing to dive into the OpenNMT toolkit ….

- Sockeye: a toolkit for NMT (newstest2017)



| Groundhog model | EN→DE | LV→EN |
|---|---|---|
| OpenNMT-lua | 19.70 | 10.53 |
| OpenNMT-py | 18.66 | 9.98 |
| Marian | 23.54 | 14.40 |
| Nematus | 23.86 | 14.32 |
| NeuralMonkey | 13.73 | 10.54 |
| Sockeye | 23.18 | 14.40 |



| Toolkit | Layers | EN→DE | LV→EN |
|---|---|---|---|
| OpenNMT-lua | 4/4 | 22.69 | 13.85 |
| OpenNMT-py | 4/4 | 21.95 | 13.55 |
| Marian | 4/4 | 25.93 | 16.19 |
| Nematus | 8/8 | 23.78 | 14.70 |
| NeuralMonkey | 1/1 | 13.73 | 10.54 |
| Sockeye | 4/4 | 25.55 | 15.92 |

# EN->DE a long time reference in WMT tasks

| | Original | Cleaned | < 100 tokens | < 80 tokens | <50 tokens |
|---|---:|---:|---:|---:|---:|
| Common Crawl | 2,399,123 | 1,947,168 | | | |
| Europarl v7 | 1,920,209 | 1,814,782 | | | |
| NewsCommentary v12 | 270,769 | 260,898 | | | |
| Rapid2016 | 1,329,041 | 109,283 | | | |
| Total | 5,919,142 | 4,132,131 | 4,115,100 | 4,059,848 | 3,577,356 |

- WMT14: Newstest 2014 20.6 (Phrase Based)
- WMT15: Newstest 2015 24.9 (Mila first NMT)
- WMT16: Newstest 2016 26.8 (Sennrich without back-translation)
  31.6 with back-translation (single system)
- WMT17: Newstest 2017 28.3 with back-translation + ensemble

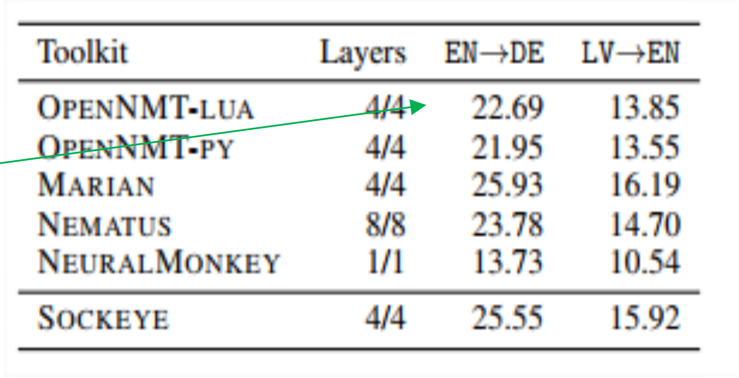
# Baseline experiment with Open NMT

- Baseline vs original ONMT paper and vs website pretrained model:
  - BRNN 2 layers of 512 + Embeddings 256
  - BPE tokenization 32K, max sequence length 100
  - Parameters: 47,714,040
  - SGD optimizer, 7 epochs, 4h10 per epoch, 29h10 training time
  - Dropout: 0.1 – Learning rate 1 during 4 epochs, then 0.5 0.2 0.1 – token batch
  - BLEU Newstest 2014: 21.84
    (original ONMT paper: 19.34 / website pretrained model:17.02)

- Same Baseline as the « Sockeye » Paper
  - 1 layer of 1024 + Embeddings 512
  - Same training schedule as above
  - BLEU Newstest 2017: 23.62 (vs 19.70 in the Sockeye Paper)

# Bigger Network experiment

|  | 2x1024+256 | 2x1024+512 | 4x1024+512 | 4x1024+256 |
|---|---|---|---|---|
| Seq Length | 100 | 100 | 80 | 80 |
| Parameters | 100,818,680 | 121,083,640 | 171,464,440 | 151,199,480 |
| Bleu NT2017 | 25.06 | 25.08 | 24.99 | 24.74 |
| Bleu NT2014 | 23.23 | 23.11 | 22.71 | 22.67 |



| Toolkit | Layers | EN→DE | LV→EN |
|---|---|---|---|
| OPENNMT-LUA | 4/4 | 22.69 | 13.85 |
| OPENNMT-PY | 4/4 | 21.95 | 13.55 |
| MARIAN | 4/4 | 25.93 | 16.19 |
| NEMATUS | 8/8 | 23.78 | 14.70 |
| NEURALMONKEY | 1/1 | 13.73 | 10.54 |
| SOCKEYE | 4/4 | 25.55 | 15.92 |

- Our experiment is much closer to the best results:
  - Marian uses a « Deep Rnn » architecture with 8 layers
  - Sockeye implemented label smoothing
- Same schedule as before, 7 epochs but we had to change the sequence length for memory constraints in the 4 layer configuration.

- Impact of sequence length

| 2x1024+256 |  |  |  |
|---|---|---|---|
| Max tokens | 100 | 80 | 50 |
| Sentences | 4,115,100 | 4,059,848 | 3,577,356 |
| Bleu NT2017 | 25.06 | 25.00 | 24.05 |
| Bleu NT2014 | 23.23 | 23.12 | 22.38 |

# Bigger Network experiment

- How do we compare to other best RNN toolkit ?

| | Newstest2014 | Newstest2015 |
|---|---|---|
| Google NMT 4 layers | 23.7 | 26.5 |
| Google NMT 8 layers | 24.4 | 27.6 |
| WMT reference | 20.6 | 24.9 |
| OpenNMT-Lua | 23.2 | 26.0 |

✓ Onmt-Lua is a bit below best tuned systems but much better than what it has been presented in some literature.

✓ Two key missing features in ONMT-Lua: Label smoothing, « GNMT attention »

# What happens with more data ?

NMT

- Rico Sennrich released the back translations used for WMT16
  - 3,579,884 additional sentences (News in-domain data)
- Our Bleu on Newstest 2016 (2x1024+512): 32.82
  - Compared to 34.2 for the best WMT16 system (ensemble + back translation)
  - Compared to 31.6 for the best single WMT16 system
- Our Bleu on Newstest 2017 (2x1024+512): 26.89
  - Compared to 28.3 for the best WMT17 Ensemble + MORE back translation (~10 M segt)
  - Compared to 26.6 for the best single WMT17 system + MORE back translation

✓ OpenNMT is highly competitive and deliver scores in line with the best WMT systems

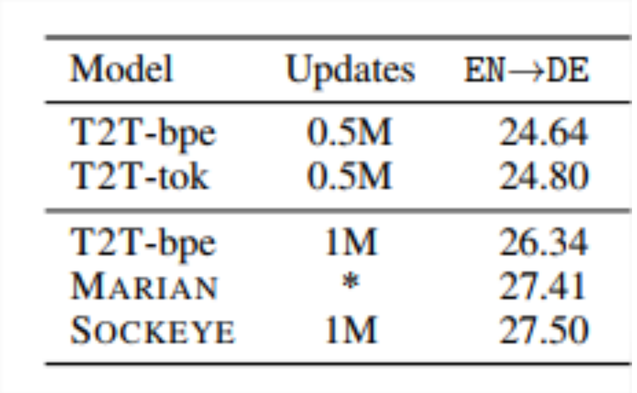
# Still State of the Art ?

- Of course not since the innovative « Transformer » from Google Brain (June 2017)
  - Feedforward Network with multi-head attention => « Attention is all you need »
  - SOA for Newstest 2014: 28.4 vs 24.6 (GNMT) vs 23.2 (ours)
- Many copycats of the Google T2T « Transformer »
  - OpenNMT-TF: very close to T2T
  - OpenNMT-Py: functional but memory issue and multi-gpu to come
  - Marian: speed and batch size not optimized – but very promising (C++)
  - Sockeye: not tested – Paper reports very good results
  - Neural Monkey: not tested
- Fairseq-py claims SOA results with Convolutional NN.

# Transformer Results

| EN-DE | Newstest2014 | Newstest2017 |
|---|---|---|
| T2T after 500k steps | 27.3 | 27.8 |
| OpenNMT-TF transformer | 26.9 | 28.0 |
| GNMT Wu (rnn) | 24.6 | |
| Onmt-Lua (rnn) | 23.2 | 25.1 |

From Sockeye Paper (newstest2017)



| Model | Updates | EN→DE |
|---|---|---|
| T2T-bpe | 0.5M | 24.64 |
| T2T-tok | 0.5M | 24.80 |
| T2T-bpe | 1M | 26.34 |
| MARIAN | * | 27.41 |
| SOCKEYE | 1M | 27.50 |

- Reminder: WMT17 best score = 28.3 with back translations
  WMT18 will be probably a Transformer with additional back translated data

✓ OpenNMT remains in the game with its TF version of the transformer

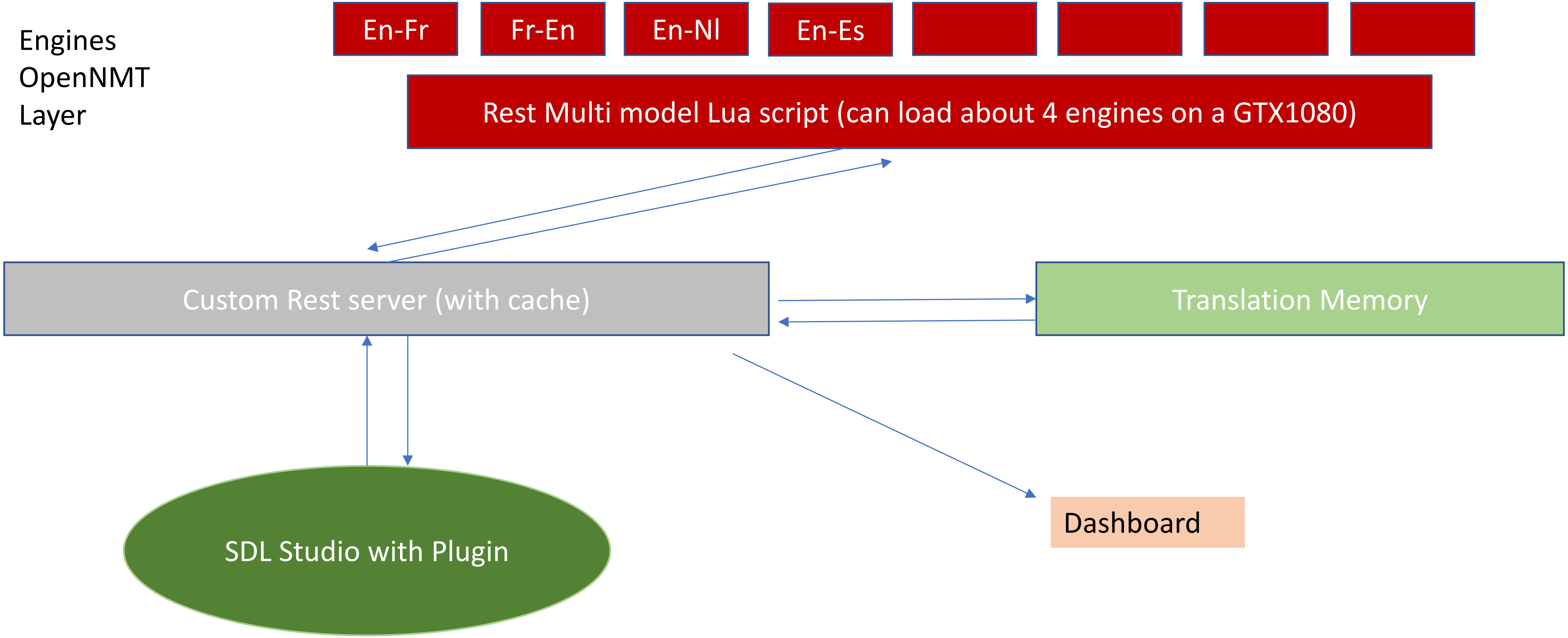
# Still State of the Art ?

- Transformer and Fairseq give much better results on WMT datasets
- Our experiments with an internal dataset (about 9 M segments) did NOT give MUCH better results
  - ONMT : Bleu 44.8
  - T2T : Bleu 45.5
  - Fairseq-py : 43.2
- Conclusion:
  - All RNN Toolkits deliver about the same performance (same technology)
  - Transformer delivers better WMT results but can be compensated by additional data and more specifically in-domain data

# Other considerations

- Translation speed
  - T2T / Onmt-Lua / Marian-nmt have similar speeds
- Model loading time
  - All tensorflow based toolkit require the models kept in memory, TF serving is not so easy to implement
  - Onmt-Lua is very fast to load
- REST API integration
  - Not the most complicated but Onmt-Lua is plug-n-play
  - Marian-Nmt provides a server as well.
- Support and community

# [Ubiqus] infrastructure



Engines
OpenNMT
Layer

En-Fr

Fr-En

En-Nl

En-Es

Rest Multi model Lua script (can load about 4 engines on a GTX1080)

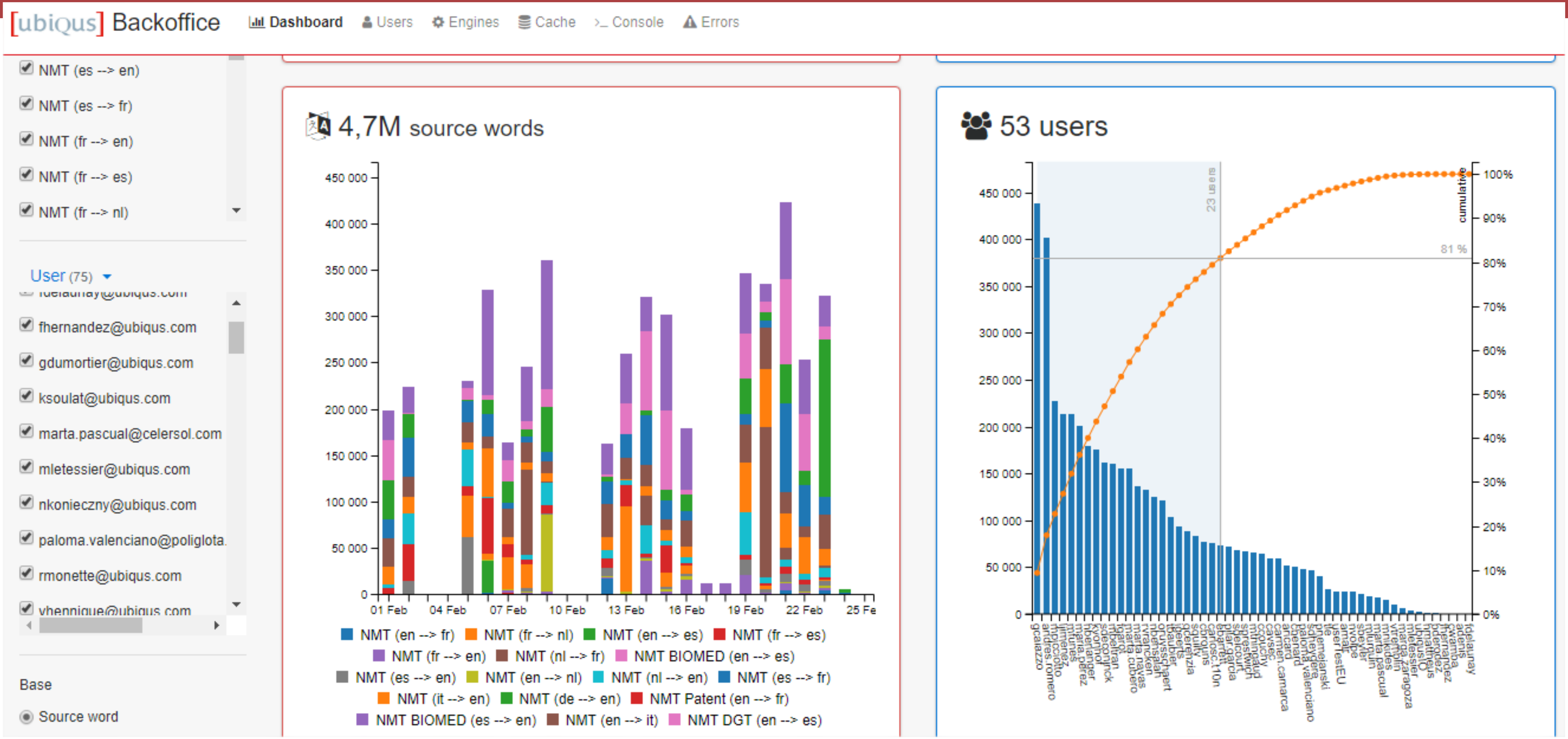
Custom Rest server (with cache)

Translation Memory

SDL Studio with Plugin

Dashboard

# Dashboard



[ubiqus] Backoffice

☑ NMT (es --> en)

☑ NMT (es --> fr)

☑ NMT (fr --> en)

☑ NMT (fr --> es)

☑ NMT (fr --> nl)

User (75)

fdelaunay@ubiqus.com

☑ fhernandez@ubiqus.com

☑ gdumortier@ubiqus.com

☑ ksoulat@ubiqus.com

☑ marta.pascual@celersol.com

☑ mletessier@ubiqus.com

☑ nkonieczny@ubiqus.com

☑ paloma.valenciano@poliglota.

☑ rmonette@ubiqus.com

☑ vhennique@ubiqus.com

Base

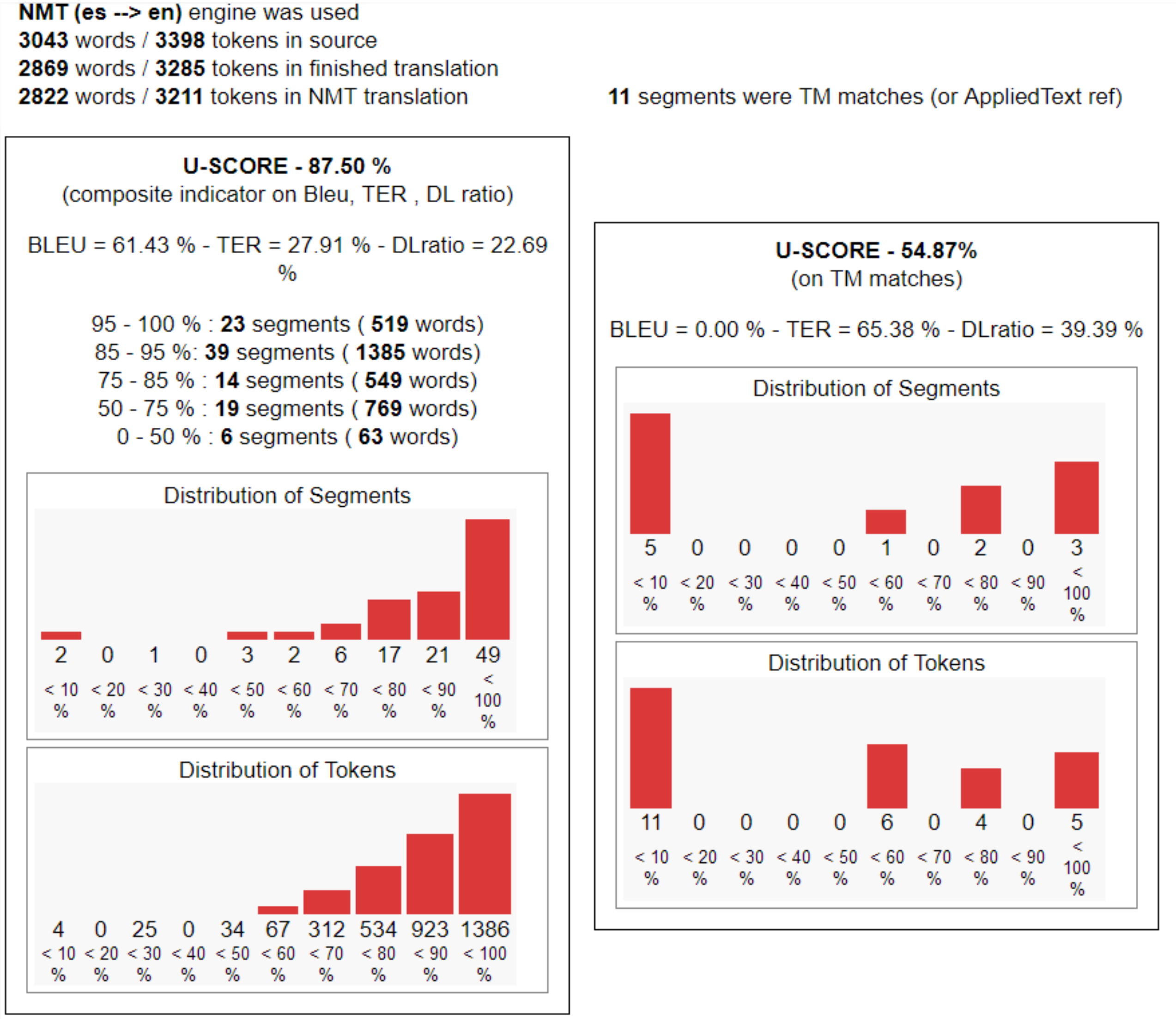
● Source word

## 4,7M source words

## 53 users

# Best way to convince is to show scores …..

- Each job sent by the human translator is scored against our NMT engine
- Gives a better sense on how accurate NMT is for each client and each type of job
- Data are then re-used for training



**NMT (es --> en)** engine was used
**3043** words / **3398** tokens in source
**2869** words / **3285** tokens in finished translation
**2822** words / **3211** tokens in NMT translation

**11** segments were TM matches (or AppliedText ref)

**U-SCORE - 87.50 %**
(composite indicator on Bleu, TER , DL ratio)

BLEU = 61.43 % - TER = 27.91 % - DLratio = 22.69 %

95 - 100 % : **23** segments ( **519** words)
85 - 95 %: **39** segments ( **1385** words)
75 - 85 % : **14** segments ( **549** words)
50 - 75 % : **19** segments ( **769** words)
0 - 50 % : **6** segments ( **63** words)

**U-SCORE - 54.87%**
(on TM matches)

BLEU = 0.00 % - TER = 65.38 % - DLratio = 39.39 %

# OpenNMT Evaluation

# Thank you !

Vincent Nguyen

OpenNMT Workshop Paris, March 2018