A Comparative Evaluation of Phrase-Based Statistical and Neural Machine Translation

TraMOOC

Translation for Massive Open Online Courses

Joss Moorkens

Translating & the Computer 38

DCU

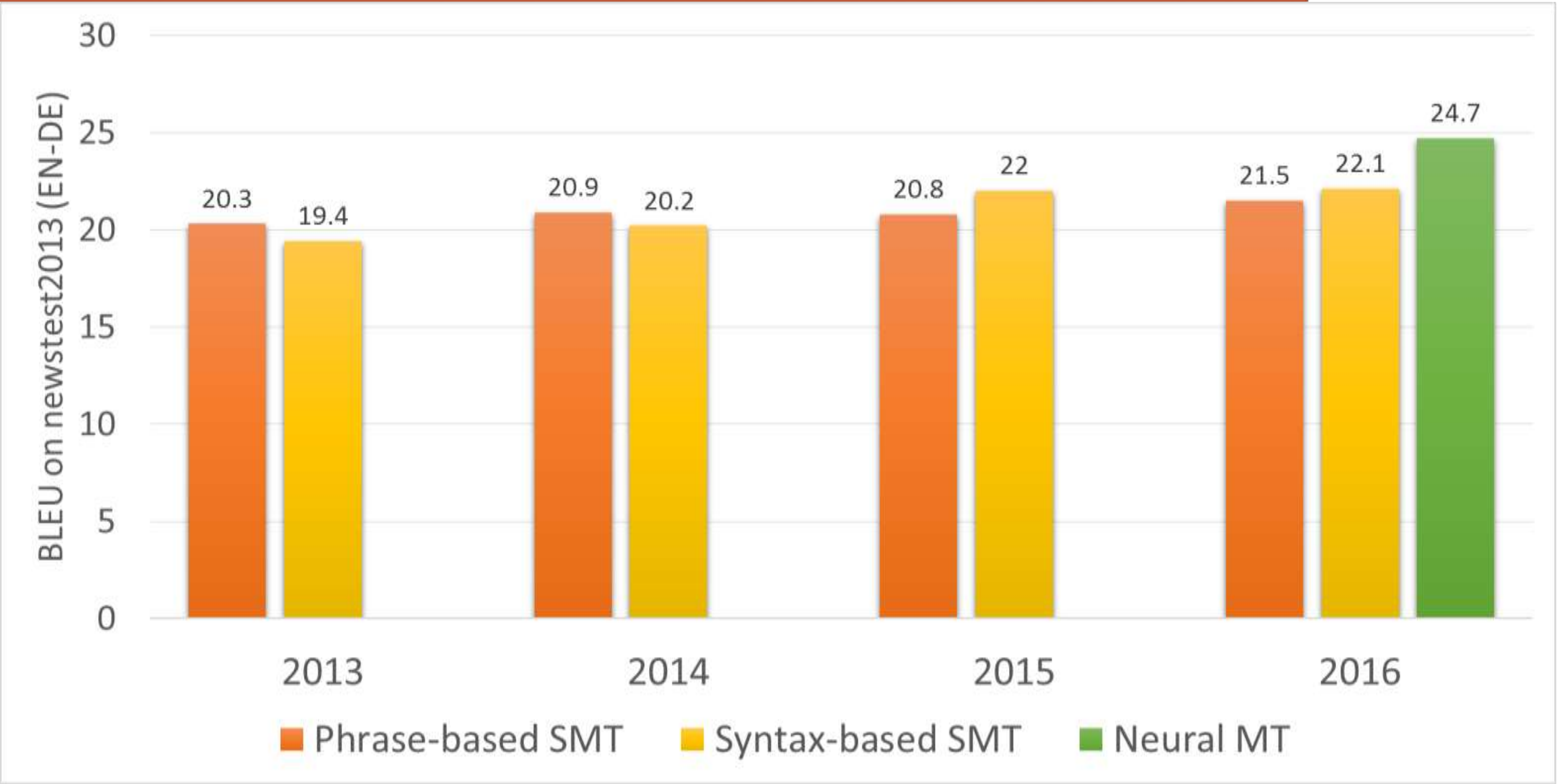
Table of contents

Co-authors: Sheila Castilho, Federico Gaspari, Andy Way, Yota Georgakopoulou, Maria Gialama, Rico Sennrich, Alexandra Birch, Antonio Miceli Barone, Valia Kordoni

Translation for Massive Open Online Courses

- **Reliable Machine Translation (MT)** for **Massive Open Online Courses (MOOCs)**
- The main expected outcome is a **high-quality semi-automated machine translation service** for educational text data on a MOOC platform
- Open educational platform for MT and a replicable process for creating such a service

What is the State of the Art for Machine Translation?



What is the State of the Art for Machine Translation?

uedin-nmt	34.2
metamind	32.3
NYU-UMontreal	30.8
cambridge	30.6
uedin-syntax	30.6
KIT/LIMSI	29.1
KIT	29.0
uedin-pbmt	28.4
jhu-syntax	26.6

EN→DE

uedin-nmt	38.6
uedin-pbmt	35.1
jhu-pbmt	34.5
uedin-syntax	34.4
KIT	33.9
jhu-syntax	31.0

DE→EN

uedin-nmt	25.8
NYU-UMontreal	23.6
jhu-pbmt	23.6
cu-chimera	21.0
uedin-cu-syntax	20.9
cu-tamchyna	20.8
cu-TectoMT	14.7
cu-mergedtrees	8.2

EN→CS

uedin-nmt	31.4
jhu-pbmt	30.4
PJATK	28.3
cu-mergedtrees	13.3

CS→EN

uedin-pbmt	35.2
uedin-nmt	33.9
uedin-syntax	33.6
jhu-pbmt	32.2
LIMSI	31.0

RO→EN

QT21-HimL-SysComb	28.9
uedin-nmt	28.1
RWTH-SYSCOMB	27.1
uedin-pbmt	26.8
uedin-lmu-hiero	25.9
KIT	25.8
lmu-cuni	24.3
LIMSI	23.9
jhu-pbmt	23.5
usfd-rescoring	23.1

EN→RO

uedin-nmt	26.0
amu-uedin	25.3
jhu-pbmt	24.0
LIMSI	23.6
AFRL-MITLL	23.5
NYU-UMontreal	23.1
AFRL-MITLL-verb-annot	20.9

EN→RU

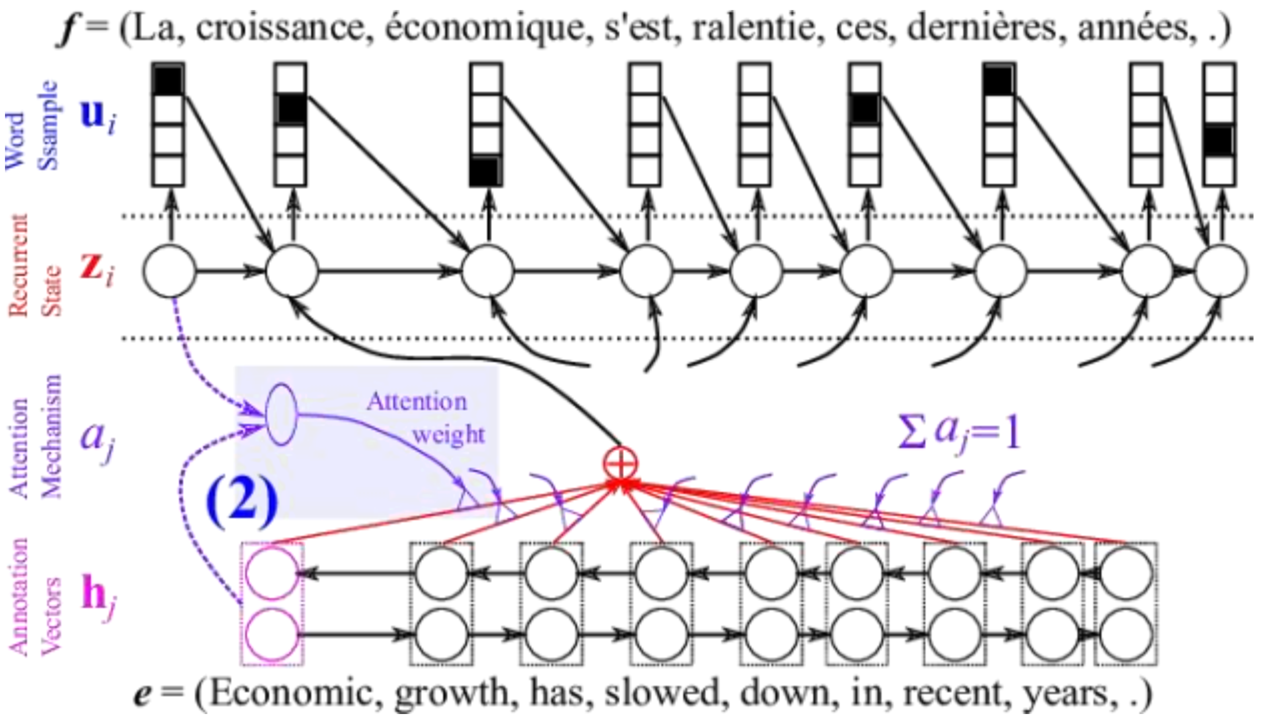
amu-uedin	29.1
NRC	29.1
uedin-nmt	28.0
AFRL-MITLL	27.6
AFRL-MITLL-contrast	27.0

RU→EN

- Edinburgh NMT
- System Combination with Edinburgh NMT

Neural Machine Translation

- SMT - many small sub-components that are tuned separately
- NMT - build and train a **single, large neural network** that reads a sentence and outputs a correct translation (Bahdanau, Cho, Bengio 2015)
- Uses a Recurrent Neural Network (RNN) to deal with variable segment length
- NMT predicts a target word based on the context associated with source and previously generated target words
- A small neural network, called an *attention mechanism* analyses context for every source word



Neural Machine Translation: Pros and Cons

- Main strength of NMT is grammatical improvements, but possible degradation in lexical transfer (Neubig, Morishita, Nakamura 2015)
- Output conditioned on full source text and target history
- End-to-end trained model
- Some problems:
  - Networks have fixed vocabulary → poor translation of rare/unknown words
  - Models are trained on parallel data; how do we use monolingual data?
  - Recent solutions:
    - Subword models allow translation of rare/unknown words (Sennrich, Birch, Haddow 2016a)
    - Train on back-translated monolingual data (Sennrich, Birch, Haddow 2016b)

Previous Studies

Neural versus Phrase-Based Machine Translation Quality: a Case Study (Bentivogli et al. 2016)

- Results show that NMT system outperforms all other approaches.
- Post-edit effort lower (-26%) on all sentence lengths
- Fewer morphology errors (-19%), lexical errors (-17%), and word order errors (-50%)
- Improvement in placement of verbs (-70% errors)

Previous Studies

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (Wu et al. 2016)

- Decreased training time and computational requirements
- Results show that NMT system strongly outperforms other approaches
- Improved translation quality for morphologically rich languages
- Human evaluation ratings closer to HT than PBSMT
- Additional tweaks required for NMT to perform well on “real data”

Our methodology

- 4 datasets (250 segments) from real EN MOOC data translated into German, Greek, Portuguese, and Russian using TraMOOC engines
- PB-SMT/NMT mixed, random task order
- 2-4 professional translators
- MT engines trained on same data: open corpora plus educational data from Coursera, Qatar Educational Domain, EU Teachers’ Corner

TraMOOC
Translation for Massive Open Online Courses

Our methodology

- Comparative ranking of 100 randomised translations
- Post-editing using PET (Aziz, Castilho, Specia 2012)
  - Temporal effort – time spent post-editing (Krings 2001)
  - Technical effort – edit count
  - Cognitive effort – pause-word-ratio (Lacruz, Denkowski, Lavie 2014)
- Rating of fluency and adequacy
- Error annotation
  - Inflectional morphology, Word order, Omission, Mistranslation, Addition

Translation for Massive Open Online Courses

Side-by-side ranking

EN-EL Evaluations	PB-SMT preference	NMT preference
400	174	226
	43.5%	56.5%
EN-DE Evaluations	**PB-SMT preference**	**NMT preference**
300	61	239
	20.3%	79.7%
EN-RU Evaluations	**PB-SMT preference**	**NMT preference**
300	110	190
	36.7%	63.3%
EN-PT Evaluations	**PB-SMT preference**	**NMT preference**
300	115	185
	38.3%	61.7%

Side-by-side ranking

NMT the preferred paradigm for all texts and language pairs

- Business Analysis UGC 65% prefer NMT
- Medical Training transcript 54% prefer NMT
- Physics transcript 52% prefer NMT
- Explaining advertising transcript 55%% prefer NMT

Short segments (20 tokens or fewer*): 53% prefer NMT

Long segments (over 20 tokens*): 61% prefer NMT

NMT: Ratings of fluency are higher

- For all 4 language pairs:

FLUENCY
1. No fluency
2. Little fluency
3. Near native
4. Native

	EN-DE		EN-EL		EN-PT		EN-RU	
% scores assigned 3-4 fluency value (SMT, NMT)	54.2	67.6	65	75	73.8	79.5	60.2	75.1
% scores assigned 1-2 fluency value (SMT, NMT)	45.8	32.4	35	25	26.2	20.5	39.8	24.9

Ratings of adequacy: mixed results

- For all 4 language pairs:

ADEQUACY
1. None of it
2. Little of it
3. Most of it
4. All of it

	EN-DE		EN-EL		EN-PT		EN-RU	
% scores assigned 3-4 adequacy value (SMT, NMT)	73.5	66.4	89	89	94.7	97.1	72.8	77.5
% scores assigned 1-2 adequacy value (SMT, NMT)	26.5	33.6	11	11	5.3	2.9	27.2	22.5

Some examples

- ST: I am just making sure that I understand this correctly.
- SMT: Estou só para ter a certeza que entendi corretamente.
    - “só” in Portuguese means ‘just’ but also “alone/lonely”
- NMT: Eu estou apenas me certificando de que eu entendo isso corretamente.

- ST: Would you send just 10 materials that are the most suitable.
- SMT: Würden Sie nur 10 Materialien, die am besten geeignet sind.
- NMT: Schicken Sie einfach 10 Materialien, die am besten geeignet sind.

Some examples

- ST: It's about copy-paste from pdf to wiki card.
- NMT: É sobre copiar-pasta de pdf para wiki card.
- SMT: Trata-se de copiar e colar de pdf para cartão wiki.

- ST: was webinar live today?
- HT: O webinar foi ao vivo hoje?
- NMT: Será que o webinar vive hoje?
- SMT: Foi webinar vivem hoje?

Post-editing: temporal effort

Words per second (all PEs)	SMT	NMT
German	0.21	0.22
Greek	0.22	0.24
Portuguese	0.29	0.30
Russian	0.14	0.14

Previous work by Moorkens & O'Brien (2015) found an average speed of 0.39 WPS for EN-DE professional PE.

SMT, NMT	German		Greek		Portuguese		Russian	
POST-EDITED SENTENCES (CHANGED)	940	813	928	863	874	844	930	848
UNCHANGED SMT, NMT	60	187	72	137	126	156	70	152

Error Markup

- Fewer overall errors for all language pairs
- Marked improvement in word order in NMT

	German		Greek		Portuguese		Russian	
	SMT	NMT	SMT	NMT	SMT	NMT	SMT	NMT
Segments with No Issues	***61***	***189***	***90***	***168***	***197***	***236***	***101***	***195***
The total number of "Inflectional morphology"	732	608	443	307	404	378	695	506
The total number of "Word Order"	382	180	303	208	216	181	197	122
The total number of "Omission"	126	84	48	57	53	58	194	163
The total number of "Addition"	46	39	24	31	61	44	183	151
The total number of "Mistranslation"	401	323	459	483	348	342	385	404
Total number of issues	**1687**	**1234**	**1277**	**1086**	**1082**	**1003**	**1654**	**1346**

Summary

- In this study, using these language pairs, in this domain…
- Fluency is improved, word order errors are fewer using NMT
- Fewer segments require editing using NMT
- NMT produces fewer morphological errors
- No clear improvement for omission or mistranslation using NMT
- NMT for production: no great improvement in post-editing throughput
    - “Errors are more difficult to spot”
- Our choice for TraMOOC: NMT

- *Still to come: analysis of technical & cognitive PE effort, initial analysis of pause duration*

TraMOOC
Translation for Massive Open Online Courses

References

Aziz, Castilho, Specia 2012 PET: a Tool for Post-editing and Assessing Machine Translation

Bahdanau, Cho, Bengio 2015 Neural Machine Translation by Jointly Learning to Align and Translate

Bentivogli et al. 2016 Neural versus Phrase-Based Machine Translation Quality: a Case Study

Krings 2001 Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes

Lacruz, Denkowski, Lavie 2014 Cognitive Demand and Cognitive Effort in Post-Editing

Neubig, Morishita, Nakamura 2015 Neural Reranking Improves Subjective Quality of Machine Translation

Moorkens, O'Brien 2015 Post-Editing Evaluations: Trade-offs between Novice and Professional Participants

Sennrich, Haddow, Birch 2016a Neural Machine Translation of Rare Words with Subword Units

Sennrich, Haddow, Birch 2016b Improving Neural Machine Translation Models with Monolingual Data

Wu et al. 2016 Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644333.