# Update on Machine Translation Research

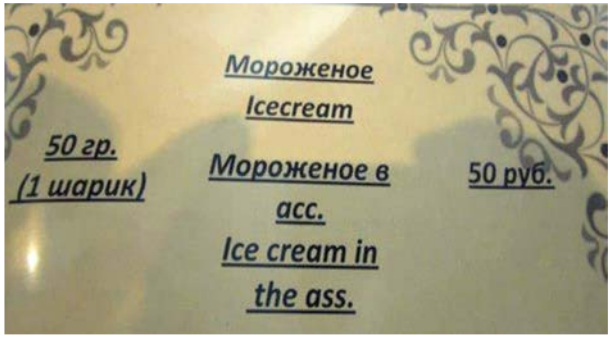
Chip Huyen (@chipro)
NVIDIA
chip@huyenchip.com

# Contents

- Milestones in machine translation
- Phrase-based vs neural MT
- Research directions in neural MT
- Evaluation and quality estimation
- What’s next

Society: “AI is putting translators out of job”

AI:





## Facebook translates 'good morning' into 'attack them', leading to arrest

**Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer**

# Machine Translation

## Timeline

Mi pyeryedayem mislyi posryedstvom ryechi

We transmit thoughts by means of speech.

**1954**

IBM 701 translated 60 Russian sentences into English

**1966**

"MT is more expensive, less accurate, slower than human translation"

- ALPAC report

**Recommendation:**
- Develop tools to aid translators

- Brown et al.

**1990**

**Statistical machine translation**

alignment, syntax-based, phrase-based

**2013**

**Neural Machine Translation**

recurrent neural networks

sequence to sequence

MSFT AI achieves human parity on WMT Chinese–English news translation (2014)

**2019+**

Document-length translation

Unsupervised systems

Low-resource languages

Quality estimation

CREATED BY

Chip Huyen (@chipro | huyenchip.com)

# Phrase-based vs Neural

# Phrase-based (PBMT) vs neural (NMT)

- Industry: “Which one is better: phrase-based MT or neural MT?”

# Phrase-based vs neural

- Industry: “Which one is better: phrase-based MT or neural MT?”
- Academia: “Why are people still doing phrase-based?”

# Compared to PBMT, NMT ...

- sounds more fluent
- makes impressively less errors: lexical errors (-16.9% on EnDe, -27.1% on EnFr), morphology errors (-31.7%, -41.4%) and word order errors (-40.9%, -48.4%)
- generates outputs that considerably lower the overall post-edit effort

Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French

# Academic benchmarks dominated by neural models

| Trend | Dataset | Best Method | Paper title |
|---|---|---|---|
|  | WMT2014 English-German | depth growing | Depth Growing for Neural Machine Translation |
|  | WMT2014 English-French | Transformer Big + BT | Understanding Back-Translation at Scale |
|  | IWSLT2015 German-English | Transformer Base + adversarial MLE | Improving Neural Language Modeling via Adversarial Training |
|  | WMT2016 English-Romanian | FlowSeq-large (NPD n = 30) | FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow |
|  | WMT2016 Romanian-English | MASS | MASS: Masked Sequence to Sequence Pre-training for Language Generation |
|  | WMT2014 German-English | FlowSeq-large (NPD n = 30) | FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow |

# Real world applications more complex than academic datasets

- Low-resource
- Long sequences
- No predefined domain

# Phrase-based vs neural

1. NMT and PBMT performances both **depend on language pairs**.

# Phrase-based vs neural

1. NMT and PBMT performances both depend on language pairs.
2. NMT systems generally **require a lot of data** (Lample et al., 2018).

# Phrase-based vs neural

1. NMT and PBMT performances both depend on language pairs.
2. NMT systems generally require a lot of data (Lample et al., 2018).
3. NMT quality **degrades more quickly as sequence length increases**. (Toral et Sanchez-Cartagena, 2017). Transformer worse than RNNs.

# Phrase-based vs neural

1. NMT and PBMT performances both depend on language pairs.
2. NMT systems generally require a lot of data (Lample et al., 2018).
3. NMT quality degrades more quickly as sequence length increases. (Toral et Sanchez-Cartagena, 2017). Transformer worse than RNNs.
4. On **multi-domain** English-French data, **PBMT outperforms** NMT (Farajian et al., 2017).

# Reduce data requirements

1. Unsupervised training
   a. Initialized with word-to-word translation, large monolingual corpora, iterative back-translation (Lample et al., 2018)

**Algorithm 1:** Unsupervised MT

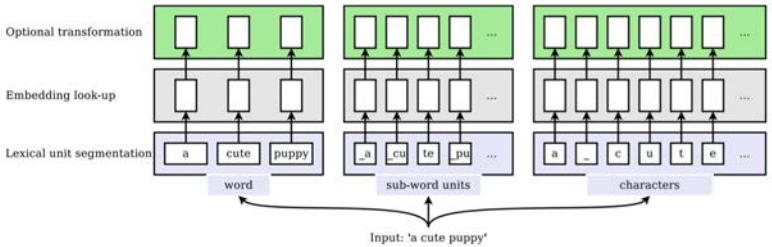
**Language models:** Learn language models $P_s$ and $P_t$ over source and target languages;
**Initial translation models:** Leveraging $P_s$ and $P_t$, learn two initial translation models, one in each direction: $P^{(0)}_{s \to t}$ and $P^{(0)}_{t \to s}$;
**for** k=1 **to** N **do**
    **Back-translation:** Generate source and target sentences using the current translation models, $P^{(k-1)}_{t \to s}$ and $P^{(k-1)}_{s \to t}$, factoring in language models, $P_s$ and $P_t$;
    Train new translation models $P^{(k)}_{s \to t}$ and $P^{(k)}_{t \to s}$ using the generated sentences and leveraging $P_s$ and $P_t$;
**end**

# Research directions in NMT

# Reduce data requirements

1. Unsupervised training
2. Multilingual translation

   split up longer words into shorter subwords to generalize across morphological variants or compounds (Wang et al., 2018)



# Reduce data requirements

1. Unsupervised training
2. Multilingual translation
3. Pretrained models

   Up to 5.3 BLEU in resource-poor setup (Edunov et al., 2019)

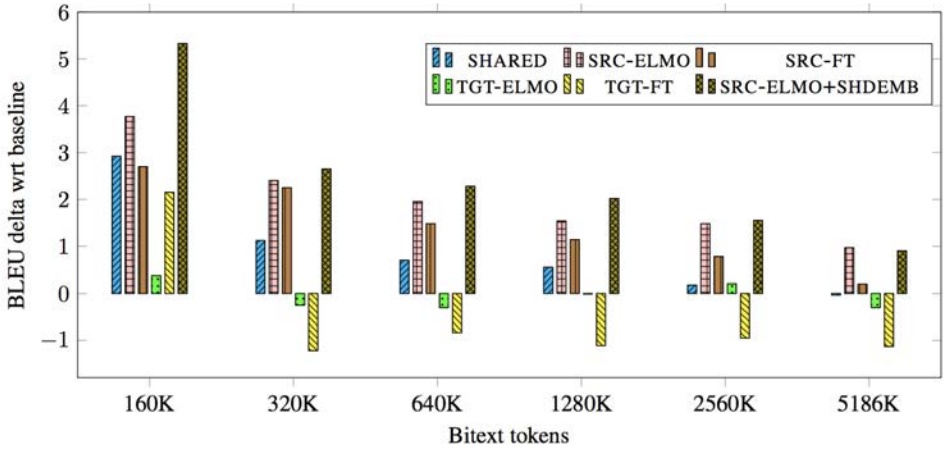
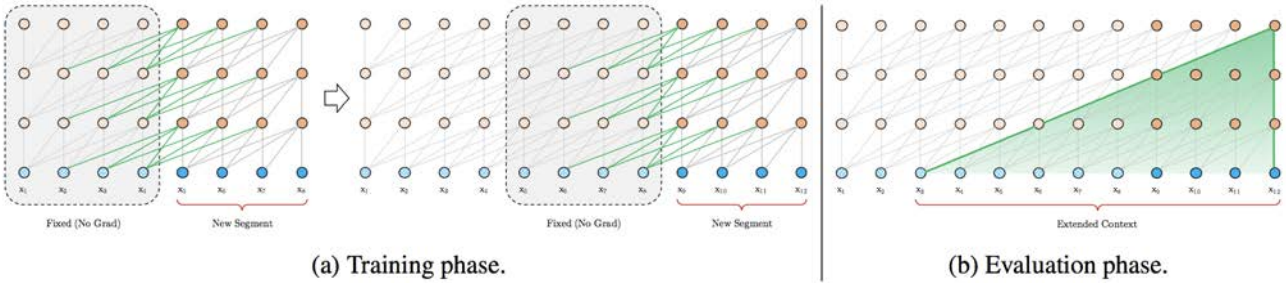



Figure 1: BLEU difference to a bitext-only baseline when adding pre-trained language model representations to a neural machine translation model in different simulated bitext settings. Results are based on averaging newstest2012-2017 of WMT English-German translation.

# Increase memory of NMT systems

1. Transformer-XL (Dai et al., 2018)
   a. previous segment state is cached and reused
   b. dependency length 450%+ compared to vanilla Transformers

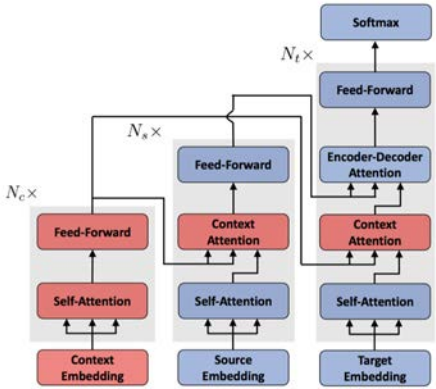


# Increase memory of NMT systems

1. Transformer-XL (Dai et al., 2018)
2. Cache-like memory network (Tu et al., 2018)
    a. stores recent hidden representations as translation history
    b. probability distribution over generated words is updated online based on translation history

# Increase memory of NMT systems

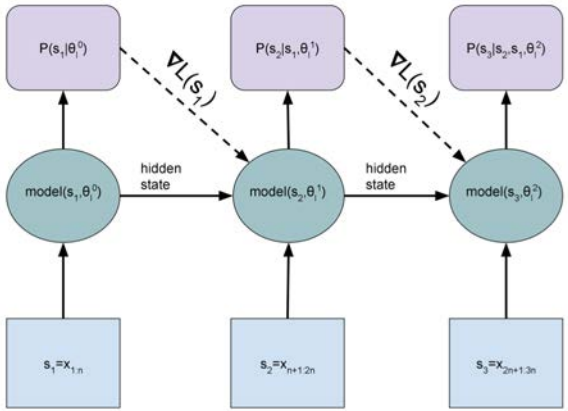
1. Transformer-XL (Dai et al., 2018)
2. Cache-like memory network (Tu et al., 2018)
3. Document-level context representation (Zhang et al., 2018)



# Domain adaptation

Dynamic evaluation (Krause et al., 2017)

- Greatly improves LMs on out of domain text
- Could be used to adapt decoder in doc or book level MT



# Neural phrase-based (NPMT)

1. locally reorders input sequences to approximate monotonic alignments with output sequences
2. explicitly models the phrase structures in output sequences
3. directly outputs phrases in a sequential order


Towards Neural Phrase-based Machine Translation (Huang et al., ICLR 2018)

# Evaluation and Quality Estimation

# Evaluation metrics

- **BLEU** and its variants **ROUGE** (recall instead of precision), **NIST** (weighted BLEU)

Example by Rachael Tatman, Evaluating Text Output in NLP: BLEU at your own risk

# BLEU

- measures **n-gram overlapping** (n <= 4)

# BLEU

- measures **n-gram overlapping** (n <= 4)

  Source: “J’ai mangé trois noisettes.”

  Generated: “I ate three hazelnuts.”

  References: “I ate three filberts.”
  “I’ve eaten three hazelnuts.”

Example by Rachael Tatman, Evaluating Text Output in NLP: BLEU at your own risk

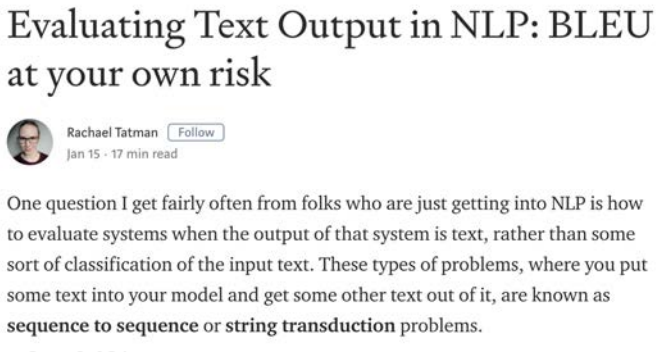
# BLEU

- measures **n-gram overlapping** (n <= 4)

  Source: “J’ai mangé trois r

  Generated: “I ate three hazeln

  References: “I ate three filberts.”
  “I’ve eaten three h



Evaluating Text Output in NLP: BLEU at your own risk

Rachael Tatman Follow
Jan 15 · 17 min read

One question I get fairly often from folks who are just getting into NLP is how to evaluate systems when the output of that system is text, rather than some sort of classification of the input text. These types of problems, where you put some text into your model and get some other text out of it, are known as **sequence to sequence** or **string transduction** problems.

Example by Rachael Tatman, Evaluating Text Output in NLP: BLEU at your own risk

# Problems with BLEU

- **Requires reference translations (ground truths)**
  - need to enumerate all possible references
  - references are not available for most real-world applications

# Problems with BLEU

- Requires reference translations (ground truths)
- **Doesn’t take into account semantics**
  - doesn’t work well with morphologically rich languages
  - “I **sold** three hazelnuts.”
    “I ate three **hazelnut**.”
    “I ate hazelnuts.”

# Problems with BLEU

- Requires reference translations (ground truths)
- Doesn’t take into account semantics
- **Doesn’t measure the amount of post-editing required**

# Problems with BLEU

- Requires reference translations (ground truths)
- Doesn’t take into account semantics
- Doesn’t measure the amount of post-editing required
- **Doesn’t map well to human judgments**

# Quality estimation

- automatically estimate the quality of machine translation output at run-time, without relying on reference translations
- shared task since 2012

Instead of evaluating translations by comparing them to reference texts, can we evaluate them by comparing them to source texts?

MEWR: Machine Translation Without Reference Text (Nguyen et Chang, 2017)

# MEWR: Machine translation Evaluation Without Reference

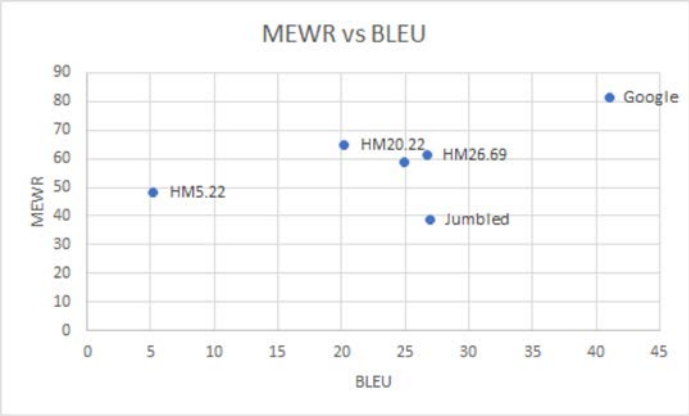
- Perceptual loss for style transfer in Computer Vision
  - evaluate the generated image by comparing its content with the content image and and its style with the style image

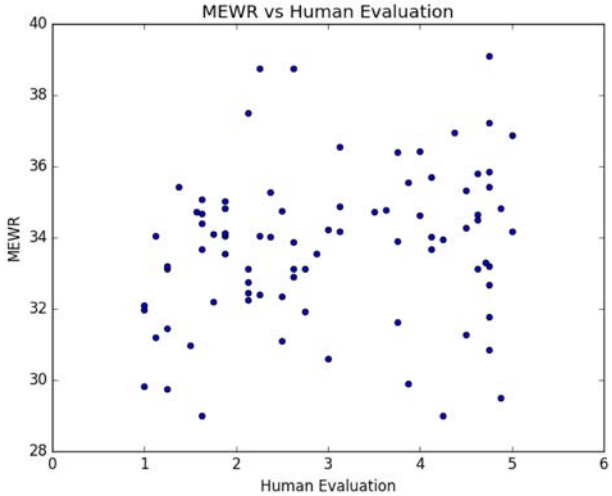
# MEWR: Machine translation Evaluation Without Reference

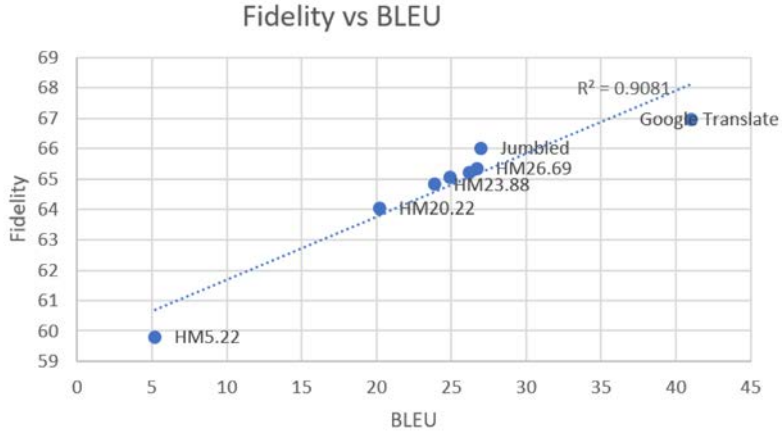
- Perceptual loss for style transfer in Computer Vision
  - evaluate the generated image by comparing its content with the content image and and its style with the style image
- Translation
  - compare style of the translation to the style of the target language, and content to the content of the source text

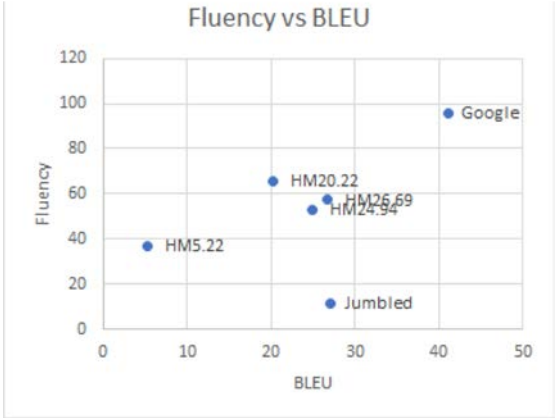
# MEWR evaluates a translation based on

- Fidelity
  - how much of the content of the source text the translation conveys
- Fluency
  - how natural it sounds to native speakers


MEWR vs BLEU
MEWR
BLEU
Google
HM20.22
HM26.69
HM5.22
Jumbled


MEWR vs Human Evaluation
MEWR
Human Evaluation


Fidelity vs BLEU
Fidelity
BLEU
R² = 0.9081
Google Translate
Jumbled
HM26.69
HM23.88
HM20.22
HM5.22


Fluency vs BLEU
Fluency
BLEU
Google
HM20.22
HM26.69
HM24.94
HM5.22
Jumbled

# What’s next

- Document-length translation
- Unsupervised/low-resource MT
- Domain adaptation
- Quality estimation
- Hybrid human-machine translation

# Thank you!

chip@huyenchip.com
Twitter: @chipro