Neural Machine Translation & Machine Translation Evaluation

Stig-Arne Grönroos

Silo.AI, stig.gronroos@silo.ai
University of Helsinki, stig-arne.gronroos@helsinki.fi

21st March 2023

About the lecturer

D.Sc. (Tech.), 2021

PostDoc, 2022–

Senior AI Scientist 2020–

Goals of the lecture

Neural machine translation

- Why NMT is the mainstream approach?
- What path led to the current state-of-the-art NMT?
- How are the current state-of-the-art NMT systems built?
- What are the challenges and limitations for the systems?

Evaluation of machine translation

- How are machine translation systems evaluated manually?
- How do the standard automatic metrics work,
  and how can they be improved?
- What are the limitations of the metrics?

Part I

Neural Machine Translation

Why neural machine translation?

Ability to generalize

Model similarity of related words and phrases

- Semantically related: synomyms, paraphrases, ...
- Morphologically related: inflections, derivations, compounds

Avoid sparsity problems encountered in phrase-based MT.

Flexibility

Different context vectors are easy to include as input.

Enables paragraph and document-level modeling.

Integration

Easier to combine with other sources of information:
Text in other languages, speech, images, videos, ...

Multitask learning

Paradigm shift to NMT

Dominant paradigm since the latter half of the 2010's.

Reasons for the paradigm shift

Increased computation power (GPUs).

Matured deep learning software frameworks and libraries: TensorFlow, (Py)Torch, Chainer, (Theano), etc.

Improvements in training algoritms for neural networks:

- Adam (Kingma and Ba 2014),
- Layer normalization (Ba, Kiros, and Hinton 2016),
- Dropout (Srivastava et al. 2014).

Cross-pollination between fields of research

Success of deep learning in computer vision and speech recognition inspired NMT.

Later, NMT architectures such as Attention and Transformers spread to other fields.

Some NMT toolkits

Fairseq

Joey NMT

Marian

OpenNMT

Sockeye

Trax

...

(Huggingface)

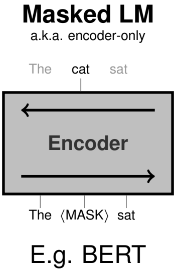
Reminder: The autoregressive language model

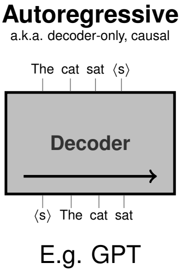
$$P(X) = \sum_i P(x_i \mid x_0, \ldots, x_{(i-1)})$$

Here is a fragment of text. Tell me how this fragment might go on.

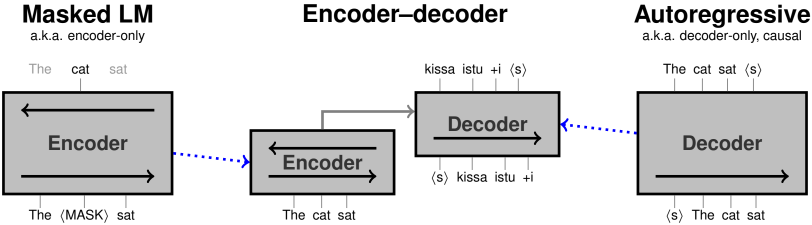
According to this model of the statistics of human language, what words are likely to come next?

Types of language model





Types of language model



E.g. BERT

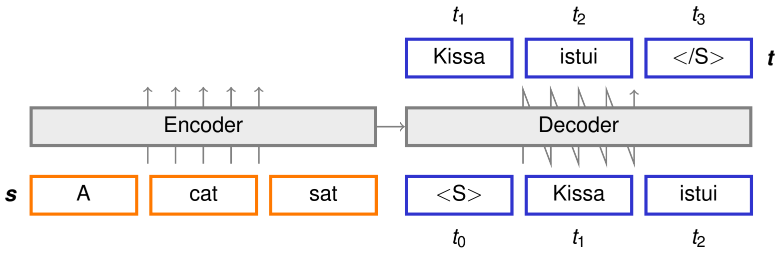
NMT models, but also e.g. T5

E.g. GPT

MT systems are conditional language models

$$P(\quad t_j \quad | \quad t_0, \ldots, t_{(j-1)}, \quad \boldsymbol{s})$$

$$P(\quad \text{“istui”} \quad | \quad \text{“<S>”, “Kissa”}, \quad \boldsymbol{s})$$



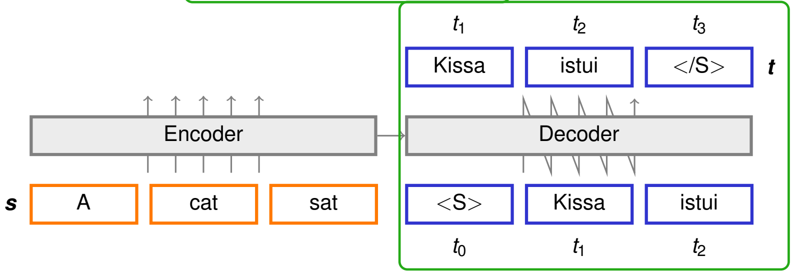
A (data-driven) MT system is a conditional language model.

Predicts the target conditioned on the source.

MT systems are conditional language models

$$P(\; t_j \;\mid\; t_0, \ldots, t_{(j-1)}, \;\boldsymbol{s})$$

$$P(\; \text{“istui”} \;\mid\; \text{“<S>”, “Kissa”}, \;\boldsymbol{s})$$



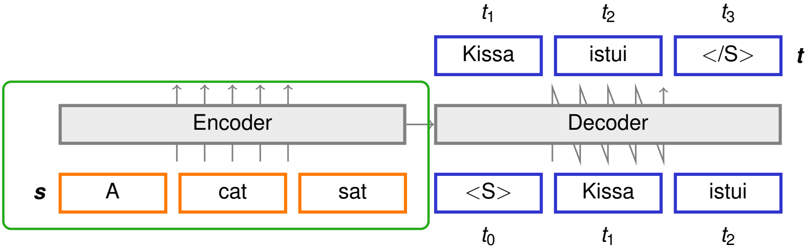
A (data-driven) MT system is a conditional language model.

Predicts the target conditioned on the source.

MT systems are conditional language models

$$P(\ t_j \ \mid\ t_0, \ldots, t_{(j-1)},\ \boxed{\boldsymbol{s}})$$

$$P(\ \text{“istui”} \ \mid\ \text{“<S>”, “Kissa”},\ \boxed{\boldsymbol{s}})$$



A (data-driven) MT system is a conditional language model.

Predicts the target conditioned on the source.

History lesson

Let's look at some of the breakthroughs leading towards current SOTA architectures, and the challenges that inspired these breakthroughs.

History: Embedding variable-length sequences

How to encode sequences (words, phrases, sentences) $x_1, x_2, \ldots, x_n$ of variable length $n \geq 1$ to fixed length representations?

History: Embedding variable-length sequences

How to encode sequences (words, phrases, sentences) $x_1, x_2, \ldots, x_n$ of variable length $n \geq 1$ to fixed length representations?

Remember from Word2vec lecture: Embeddings such as word2vec give fixed-length vectors for each of the units in the sequence, but the sequence itself is variable-length.

History: Embedding variable-length sequences

How to encode sequences (words, phrases, sentences) $x_1, x_2, \ldots, x_n$ of variable length $n \geq 1$ to fixed length representations?

Remember from Word2vec lecture: Embeddings such as word2vec give fixed-length vectors for each of the units in the sequence, but the sequence itself is variable-length.

But how to combine them? Summing or averaging discards the sequence order.

History: Embedding variable-length sequences

How to encode sequences (words, phrases, sentences) $x_1, x_2, \ldots, x_n$ of variable length $n \geq 1$ to fixed length representations?

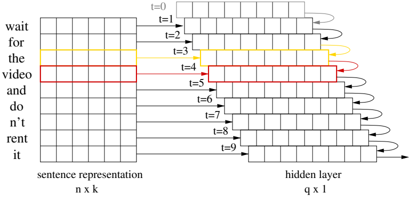
Remember from Word2vec lecture: Embeddings such as word2vec give fixed-length vectors for each of the units in the sequence, but the sequence itself is variable-length.

But how to combine them? Summing or averaging discards the sequence order.

Remember from LM Lecture: Neural network language models are able to store information over long contexts.

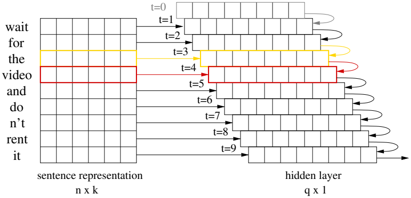
History: Sequence encoding with RNN and CNN

Recurrent neural networks: Take the last hidden state as sentence embedding. (Elman 1990; Mikolov et al. 2010)

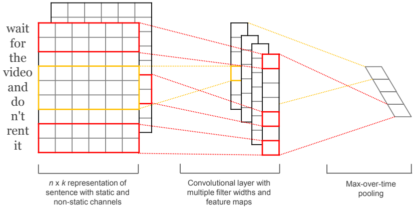


History: Sequence encoding with RNN and CNN

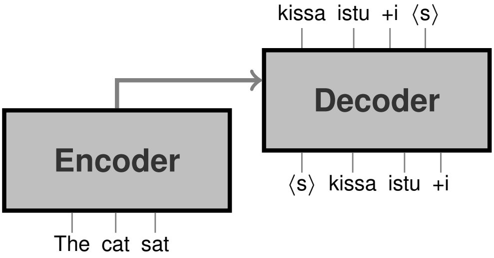
Recurrent neural networks: Take the last hidden state as sentence embedding. (Elman 1990; Mikolov et al. 2010)



Alternative: Convolutional neural networks (Fukushima 1980; Kim 2014)



History: Encoder-decoder model



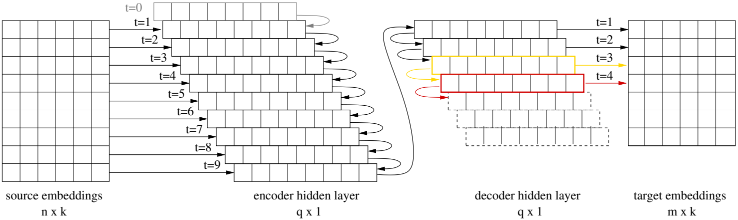
History: Sequence decoding

How to implement the decoder?

History: Sequence decoding

How to implement the decoder?

Again, we can use an RNN language model — just initialize the hidden state with the sentence representation from encoder!

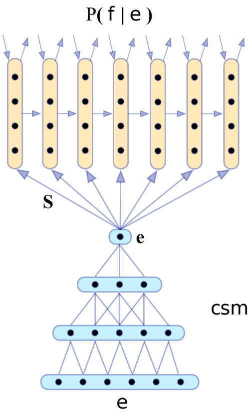


History: First complete NMT systems

Kalchbrenner and Blunsom 2013: Encode with convolutional neural networks (CNN), decode with recurrent neural network (RNN) language model

Sutskever, Vinyals, and Le 2014: Encode and decoder with RNN with long short-term memory (LSTM) units

Cho et al. 2014b: Encode with RNN with gated recursive units (GRU) or gated recursive CNN, decoder with RNN with GRUs

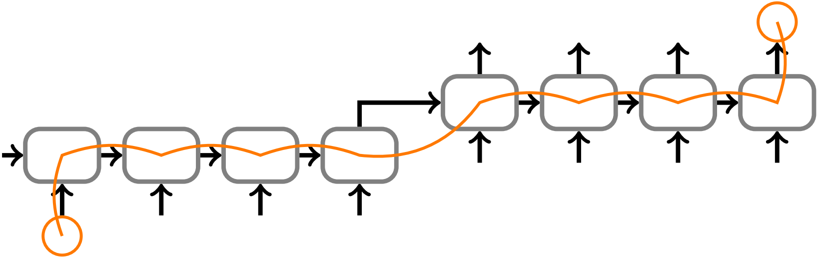


RCTM I

History: Vanishing gradient problem

The error signal decreases exponentially with the number of layers in backpropagation and gradient-based learning.

The RNN encoder must process entire sentence before sentence encoding is ready: The long path makes it hard to learn relevant information from first time steps (beginning of sentence).



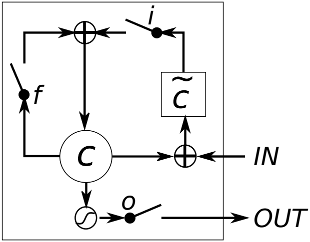
History: Gated units in recurrent neural networks

Solution:

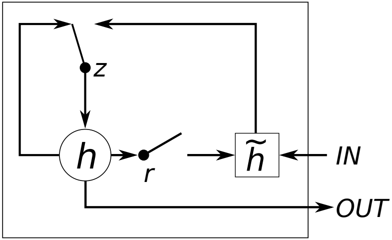
- Predict what information to keep and what to forget from the state representation.
- Gates: sigmoid activation (0–1) followed by pointwise multiplication with the target signal.

History: Gated units in recurrent neural networks

LSTM and GRU are two gate architectures with similar performance (Chung et al. 2014)



Long short-term memory
(Hochreiter and Schmidhuber 1997)



Gated recurrent unit
(Cho et al. 2014a)

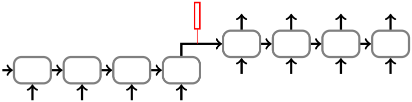
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

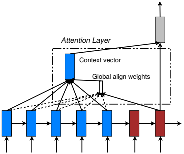
History: Attention model

Even with gated units, it is hard to decode a sensible target sentence from a single embedded source vector.

Encoder provides embeddings for each input unit — allow decoder to look at them.

Attention model: At each decoder time step, predict which parts of the source encoding are relevant for next output.

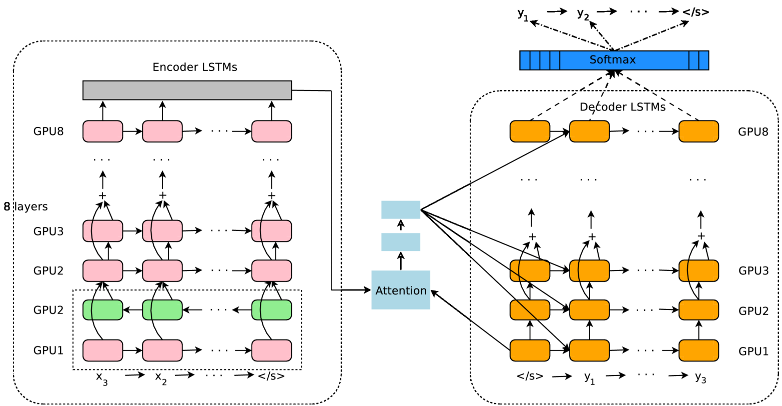




(Luong, Pham, and Manning 2015)
(Bahdanau, Cho, and Bengio 2015)

http://distill.pub/2016/augmented-rnns/#attentional-interfaces

History: Adding layers



Google NMT (Wu et al. 2016)

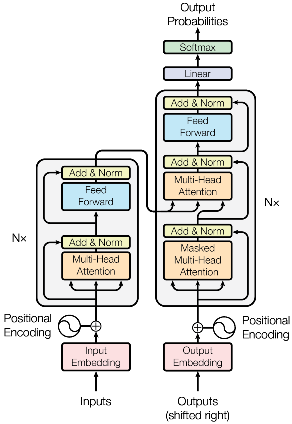
History: Transformer architecture

Recurrent networks require sequential computation ($O(n)$ for n units in sentence)

Can we cope without them?

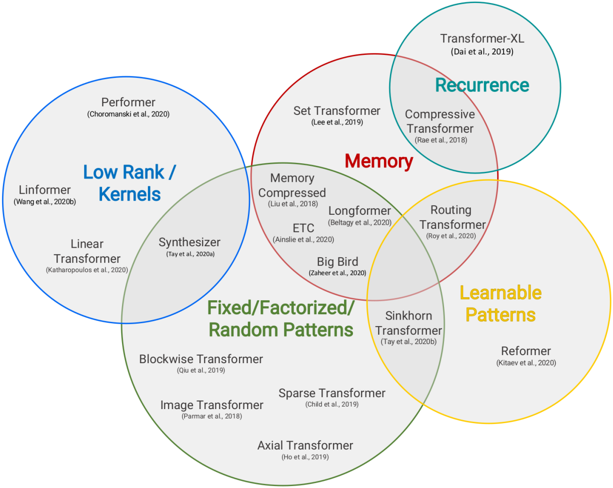
"Attention is all you need" — Google's Transformer architecture (Vaswani et al. 2017)

Multiple layers of attention networks in both encoder and decoder



https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformers: the sequels



Taxonomy of efficient Transformer architectures (Tay et al. 2020).

Mixture of Experts

In Mixture of Experts (MoE), a gating network selects which subnetworks to use for the example.

- A sparse network: not all parameters are used each time.
- Can be combined with the Transformer architecture.

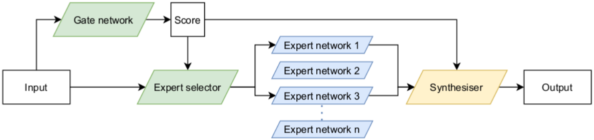


Figure 1: An illustrative example of an MoE layer. In this example, expert 1 and expert 3 are selected by the gate for computation.

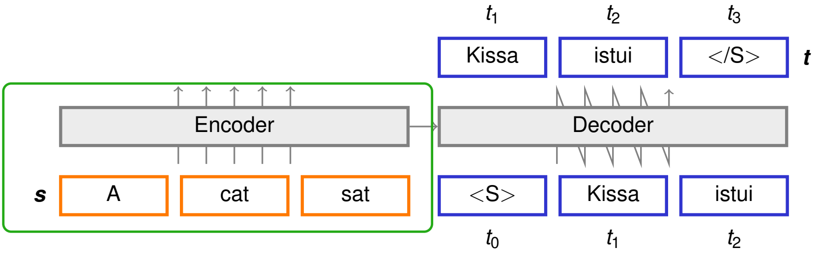
- Figure from (He et al. 2021)

Discuss in groups (a few minutes)

What happens if the conditioning on the source fails to be learned?

$$P(t_j \mid t_0, \ldots, t_{(j-1)}, \boxed{\boldsymbol{s}})$$

$$P(\text{“istui”} \mid \text{“<S>”}, \text{“Kissa”}, \boxed{\boldsymbol{s}})$$



Is Transformer all you need?

At the moment, Transformer is the state-of-the-art and *de facto* standard in NMT.

Is Transformer all you need?

At the moment, Transformer is the state-of-the-art and *de facto* standard in NMT.

But the model architecture is not everything!

Is Transformer all you need?

At the moment, Transformer is the state-of-the-art and *de facto* standard in NMT.

But the model architecture is not everything!

Especially for low-resource language pairs and morphologically rich languages, we need methods for:

Is Transformer all you need?

At the moment, Transformer is the state-of-the-art and *de facto* standard in NMT.

But the model architecture is not everything!

Especially for low-resource language pairs and morphologically rich languages, we need methods for:

1. Learning from bilingual data in other languages
2. Using monolingual corpora in source or target language
3. Selecting input and output units

Transfer learning

Current machine learning methods are data-hungry.

The easiest way to improve performance is to train on larger data.

Transfer learning

Current machine learning methods are data-hungry.

The easiest way to improve performance is to train on larger data.

Either collect more data for the task (expensive!),

Transfer learning

Current machine learning methods are data-hungry.

The easiest way to improve performance is to train on larger data.

Either collect more data for the task (expensive!),
or figure out a way to use existing data sets.

Transfer learning

Current machine learning methods are data-hungry.

The easiest way to improve performance is to train on larger data.

Either collect more data for the task (expensive!),

or figure out a way to use existing data sets.

- Labeled data for other tasks.
- Unlabeled data.

Labeled and unlabeled in the context of MT

Let's say the goal is a English-to-Finnish system.

Labeled data for this task: English-Finnish sentence pairs

- Input English sentence
- is labeled by output Finnish sentence.

Labeled and unlabeled in the context of MT

Let's say the goal is a English-to-Finnish system.

Labeled data for this task: English-Finnish sentence pairs

- Input English sentence
- is labeled by output Finnish sentence.

Labeled data for another task:

- e.g. English-Estonian sentence pairs.

Labeled and unlabeled in the context of MT

Let's say the goal is a English-to-Finnish system.

Labeled data for this task: English-Finnish sentence pairs

- Input English sentence
- is labeled by output Finnish sentence.

Labeled data for another task:

- e.g. English-Estonian sentence pairs.

Unlabeled data:

- Monolingual English,
- or monolingual Finnish.

Transfer learning techniques

Transfer learning: Use knowledge gained from solving one task in a related task.

How are the different learning tasks timed?

- Sequential transfer
- Parallel transfer
- Mix: Scheduled multi-task learning

Transfer learning techniques

Sequential transfer

Parallel transfer

Mix: Scheduled multi-task learning

Transfer learning techniques

Sequential transfer

- Often called just "transfer learning"

1. Train a system on one task (“pretraining”),
2. then transfer the knowledge,
3. and finally continue training on another task (“fine-tuning”).

Parallel transfer

Mix: Scheduled multi-task learning

Transfer learning techniques

Sequential transfer

- Often called just "transfer learning"

1. Train a system on one task (“pretraining”),
2. then transfer the knowledge,
3. and finally continue training on another task (“fine-tuning”).

- Risk: catastrophic forgetting
- Benefit: the second task doesn’t need to be known when training the first!

Parallel transfer

Mix: Scheduled multi-task learning

Transfer learning techniques

Sequential transfer

- Often called just "transfer learning"

1. Train a system on one task (“pretraining”),
2. then transfer the knowledge,
3. and finally continue training on another task (“fine-tuning”).

- Risk: catastrophic forgetting
- Benefit: the second task doesn’t need to be known when training the first!

Parallel transfer

- Often called "multi-task learning"
- Learn multiple related tasks at the same time.

Mix: Scheduled multi-task learning

Transfer learning techniques

Sequential transfer

- Often called just "transfer learning"

1. Train a system on one task (“pretraining”),
2. then transfer the knowledge,
3. and finally continue training on another task (“fine-tuning”).

- Risk: catastrophic forgetting
- Benefit: the second task doesn’t need to be known when training the first!

Parallel transfer

- Often called "multi-task learning"
- Learn multiple related tasks at the same time.

Mix: Scheduled multi-task learning

- e.g. multi-task pretraining + multi-task fine-tuning

Cross-lingual transfer: Settings

Given training data between languages A and B, can it help translating from language C to D?

Training a multilingual MT system is a multi-task training scenario

- Each language pair is one task.

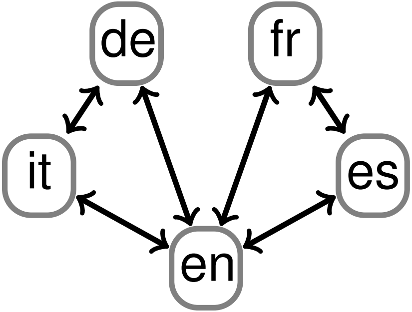
Multilingual settings:

- one-to-many
- many-to-one
- many-to-many

(Can also combine with monolingual tasks).

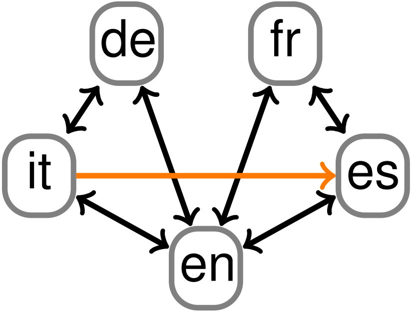
Cross-lingual transfer: Zero-shot and universal translation

Many-to-many translation enables new language pairs without training data ("zero-shot") or explicit pivot language.



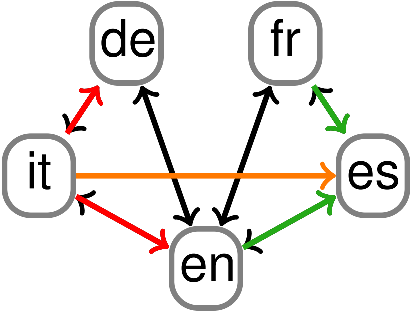
Cross-lingual transfer: Zero-shot and universal translation

Many-to-many translation enables new language pairs without training data ("zero-shot") or explicit pivot language.



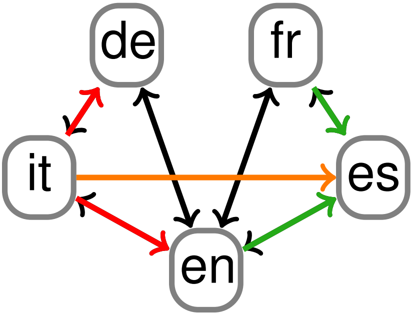
Cross-lingual transfer: Zero-shot and universal translation

Many-to-many translation enables new language pairs without training data ("zero-shot") or explicit pivot language.



Cross-lingual transfer: Zero-shot and universal translation

Many-to-many translation enables new language pairs without training data ("zero-shot") or explicit pivot language.



Universal translation: Extension of many-to-many translation to cover all languages.

How effective is cross-lingual transfer?

When model capacity is insufficient and tasks are too different, you get interference (negative transfer).

The accepted wisdom used to be that cross-lingual transfer

- is good for medium and low-resource languages,
- but for high-resource pairs bilingual was better.

Recently this was put in question

- Facebook AI's WMT 2021 News task submission (Tran et al. 2021)
- Large enough multilingual models outperform single-pair models even for high-resource language pairs like En$\leftrightarrow$Cs, En$\leftrightarrow$De, En$\leftrightarrow$Ru.

Massively multilingual machine translation

No Language Left Behind: Scaling Human-Centered Machine Translation (Costa-jussà et al. 2022)

- Massively multilingual model from Meta AI.
- Supports 200 languages,
  including both high- and low-resource languages.
- Transformer Mixture-of-Experts (MoE),
  54.5B parameters in total.

Using monolingual corpora

There is no separate language model component in NMT.

Using monolingual corpora

There is no separate language model component in NMT.

How to exploit abundant monolingual data?

Using monolingual corpora

There is no separate language model component in NMT.

How to exploit abundant monolingual data?

Approaches:

- Pretraining
- Autoencoding
- Back-translation

Monolingual corpora: Pretraining

Sequential transfer: Train a component of the model on monolingual data.

Monolingual corpora: Pretraining

Sequential transfer: Train a component of the model on monolingual data.

1. Pretrained source or target embeddings

Monolingual corpora: Pretraining

Sequential transfer: Train a component of the model on monolingual data.

1. Pretrained source or target embeddings
2. Pretrained subnetwork (encoder or decoder)

Monolingual corpora: Pretraining

Sequential transfer: Train a component of the model on monolingual data.

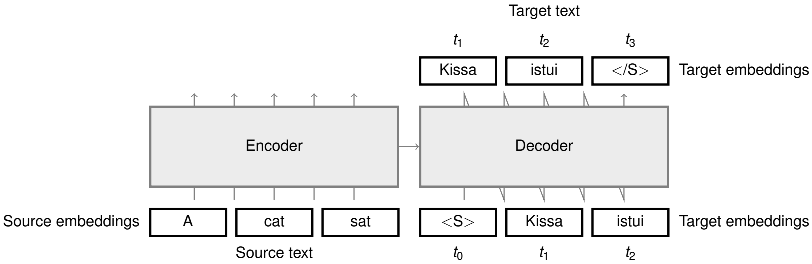
1. Pretrained source or target embeddings
2. Pretrained subnetwork (encoder or decoder)
3. Pretraining entire network (encoder and decoder)
    - E.g. finetuning a multilingual LM for MT (Liu et al. 2020)

Monolingual corpora: Pretraining

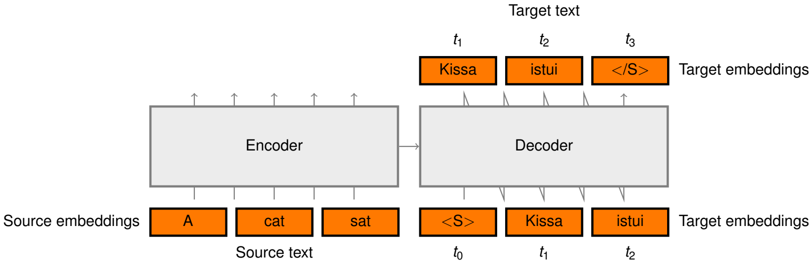
Sequential transfer: Train a component of the model on monolingual data.

1. Pretrained source or target embeddings
2. Pretrained subnetwork (encoder or decoder)
3. Pretraining entire network (encoder and decoder)
   - E.g. finetuning a multilingual LM for MT (Liu et al. 2020)
4. Language model fusion

Parameter sharing in NMT transfer learning

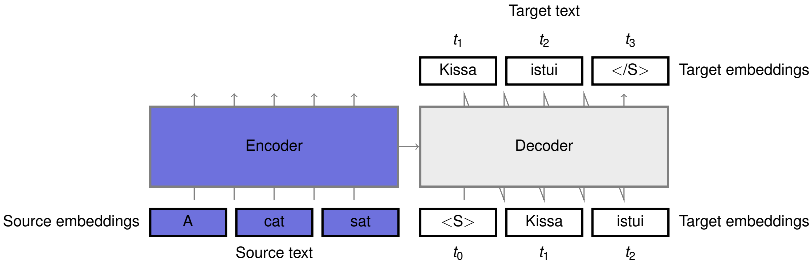


Parameter sharing in NMT transfer learning



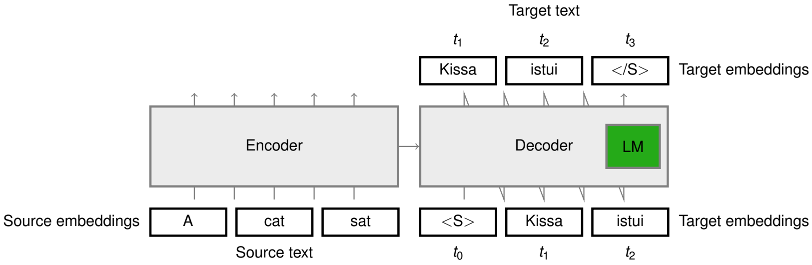
Pretrained embeddings

Parameter sharing in NMT transfer learning



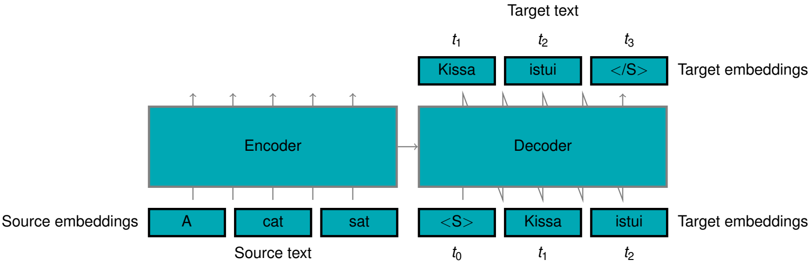
Pretrained encoder

Parameter sharing in NMT transfer learning



Language model fusion

Parameter sharing in NMT transfer learning



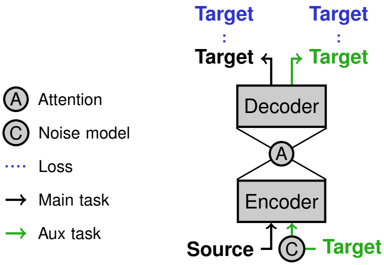
Full parameter sharing

Monolingual corpora: Autoencoding

Parallel transfer: Use multi-task learning with source-to-source or target-to-target autoencoding as an additional task.

Monolingual corpora: Autoencoding

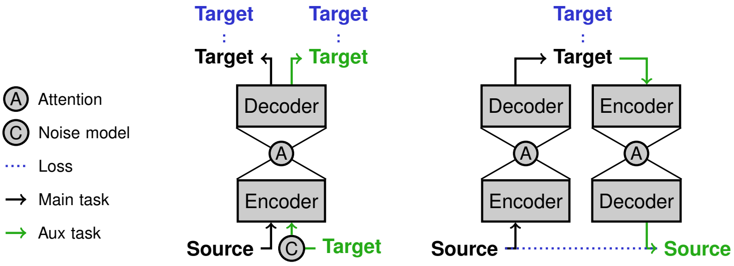
Parallel transfer: Use multi-task learning with source-to-source or target-to-target autoencoding as an additional task.



Denoising autoencoder
for target language

Monolingual corpora: Autoencoding

Parallel transfer: Use multi-task learning with source-to-source or target-to-target autoencoding as an additional task.



Denoising autoencoder for target language

Dual learning (Xia et al. 2016)

Monolingual corpora: Back-translation

Let's say the goal is a English-to-Finnish system.

Monolingual corpora: Back-translation

Let's say the goal is a English-to-Finnish system.

First train a Finnish-to-English system and translate any monolingual Finnish corpora with it.

Monolingual corpora: Back-translation

Let's say the goal is a English-to-Finnish system.

First train a Finnish-to-English system and translate any monolingual Finnish corpora with it.

Use results as additional training material.

- Synthetic training data.

Monolingual corpora: Back-translation

Let's say the goal is a English-to-Finnish system.

First train a Finnish-to-English system and translate any monolingual Finnish corpora with it.

Use results as additional training material.

- Synthetic training data.

This technique is called back-translation (Sennrich, Haddow, and Birch 2016a).

Monolingual corpora: Back-translation

Let's say the goal is a English-to-Finnish system.

First train a Finnish-to-English system and translate any monolingual Finnish corpora with it.

Use results as additional training material.

- Synthetic training data.

This technique is called back-translation (Sennrich, Haddow, and Birch 2016a).

Bad translations on the source side do not matter too much.

Monolingual corpora: Back-translation

Let's say the goal is a English-to-Finnish system.

First train a Finnish-to-English system and translate any monolingual Finnish corpora with it.

Use results as additional training material.

- Synthetic training data.

This technique is called back-translation (Sennrich, Haddow, and Birch 2016a).

Bad translations on the source side do not matter too much.

Large gains, but double work in training.

Lexical units in NMT

Limiting issues in phrase-based MT:

- Many tokens per sentence makes decoding more difficult.
- Different number of tokens in source and target sentence makes word alignment more difficult.

No such restrictions in NMT!

Units for encoder and decoder

Encoder input symbols

Words: large vocabulary, rare words, OOVs.

- but factors (e.g. morphological analysis) easy to integrate.

Using characters may slow down attention too much.

- Softmax operation on input tokens.

Decoder output symbols

Computational complexity increases with vocabulary size due to softmax in output layer.

Conclusion

Subword units (morphological segmentation if available, or statistical subwords) may be a good compromise.

Multilingual units

Current standard practice in segmentation:

- SentencePiece (Kudo 2018)

Still popular:

- Byte-pair encoding (BPE) (Sennrich, Haddow, and Birch 2016b)

Joint segmentation: The source and target language corpora — or more languages in a multilingual system — can be combined as a single training corpus for SentencePiece / BPE.

- Very practical for massively multilingual models: no need for language-specific preprocessing.

Challenges

Training SOTA-size models is computationally expensive.

- Increasing the number of layers improves results but requires more GPU/TPU resources.
- Distributed training over enormous number of GPUs.

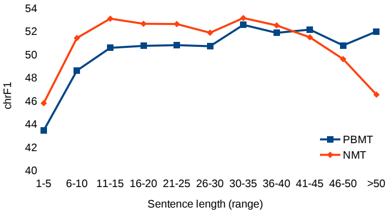
NMT is a "black box" system.

- No "phrase table" to observe or modify.
- Inconvinient especially for translation industry, where correct terminology is very important.

Challenges (cont.)

Translation quality issues

- Problems with long texts.
  - Long sentences used to be problematic (Toral and Sánchez-Cartagena 2017)

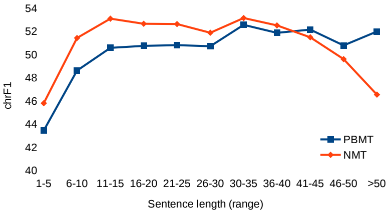


  - Now the challenge is in document-level translation.
- Good fluency, but sometimes very misleading translations — can be less predictable than PBMT

Challenges (cont.)

Translation quality issues

- Problems with long texts.
  - Long sentences used to be problematic (Toral and Sánchez-Cartagena 2017)

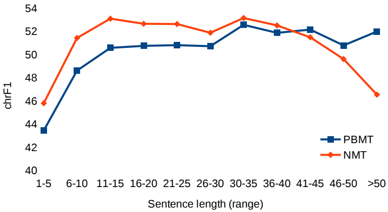


  - Now the challenge is in document-level translation.
- Good fluency, but sometimes very misleading translations — can be less predictable than PBMT
  - EN: Stealing food is a common crime in student halls.

Challenges (cont.)

Translation quality issues

- Problems with long texts.
  - Long sentences used to be problematic (Toral and Sánchez-Cartagena 2017)



  - Now the challenge is in document-level translation.
- Good fluency, but sometimes very misleading translations — can be less predictable than PBMT
  - EN: Stealing food is a common crime in student halls.
    FI: Lapsenteko on yhteistä rikollisuutta.
    (Making children is shared crime.)

Bibliography

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980. URL: http://arxiv.org/abs/1412.6980.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer normalization". In: *arXiv preprint arXiv:1607.06450*. URL: https://arxiv.org/abs/1607.06450.

Srivastava, Nitish et al. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15, pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Elman, Jeffrey L. (Mar. 1990). "Finding Structure in Time". en. In: *Cognitive Science* 14.2, pp. 179–211. ISSN: 03640213.

Bibliography (cont.)

Mikolov, Tomáš et al. (2010). "Recurrent Neural Network Based Language Model". en. In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 1045–1048.

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4, pp. 193–202.

Kim, Yoon (2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1746–1751. URL: `http://www.aclweb.org/anthology/D14-1181`.

Bibliography (cont.)

Kalchbrenner, Nal and Phil Blunsom (2013). "Recurrent Continuous Translation Models". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1700–1709. URL: http://www.aclweb.org/anthology/D13-1176.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. V Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 3104–3112. URL: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Cho, Kyunghyun et al. (2014b). "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, pp. 103–111. URL: http://www.aclweb.org/anthology/W14-4012.

Bibliography (cont.)

Chung, Junyoung et al. (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *CoRR* abs/1412.3555. URL: http://arxiv.org/abs/1412.3555.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-term Memory". In: *Neural Comput.* 9.9, pp. 1735–1780. ISSN: 0899-7667. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Cho, Kyunghyun et al. (2014a). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. URL: http://www.aclweb.org/anthology/D14-1179.

Bibliography (cont.)

Luong, Thang, Hieu Pham, and Christopher D. Manning (2015). “Effective Approaches to Attention-based Neural Machine Translation”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421. URL: http://aclweb.org/anthology/D15-1166.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *ICLR*. URL: http://arxiv.org/pdf/1409.0473v6.pdf.

Wu, Yonghui et al. (2016). “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144. URL: http://arxiv.org/abs/1609.08144.pdf.

Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Bibliography (cont.)

Tay, Yi et al. (2020). *Efficient Transformers: A Survey*. arXiv: 2009.06732 [cs.LG].

He, Jiaao et al. (2021). “FastMoE: A fast mixture-of-expert training system”. In: *arXiv preprint arXiv:2103.13262*.

Tran, Chau et al. (2021). “Facebook AI’s WMT21 News Translation Task Submission”. In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 205–215. URL: https://aclanthology.org/2021.wmt-1.19.

Costa-jussà, Marta R et al. (2022). “No language left behind: Scaling human-centered machine translation”. In: *arXiv preprint arXiv:2207.04672*.

Liu, Yinhan et al. (2020). “Multilingual denoising pre-training for neural machine translation”. In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742.

Bibliography (cont.)

Xia, Yingce et al. (Nov. 2016). "Dual Learning for Machine Translation". en. In: *arXiv:1611.00179 [cs]*. arXiv: 1611.00179. URL: http://arxiv.org/abs/1611.00179.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016a). "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 86–96. URL: http://www.aclweb.org/anthology/P16-1009.

Kudo, Taku (2018). "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". en. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 66–75. URL: http://arxiv.org/abs/1804.10959.

Bibliography (cont.)

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016b). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. URL: http://www.aclweb.org/anthology/P16-1162.

Toral, Antonio and Víctor M. Sánchez-Cartagena (2017). "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions". In: *CoRR* abs/1701.02901. URL: http://arxiv.org/abs/1701.02901.

Part II

Machine Translation Evaluation

Outline

Human evaluation

Automatic evaluation

Meta-evaluation

How to evaluate MT systems?

Final evaluation should depend on the intended application

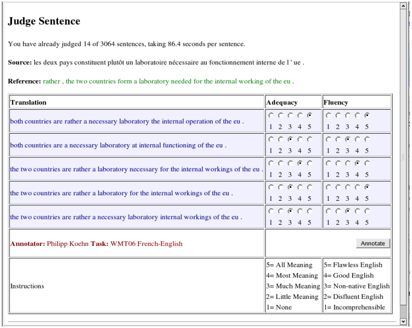
- Understanding text as it is; skimming/gisting → Human evaluation
- Aid for human translations → Decrease in translation time
- Multilingual information retrieval → IR evaluation

Human evaluation: Direct assessment

Given translation output and source and/or reference translation, how good the translation is?

Adequacy: Does the output convey the same meaning?

Fluency: Is the output good and fluent language?



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

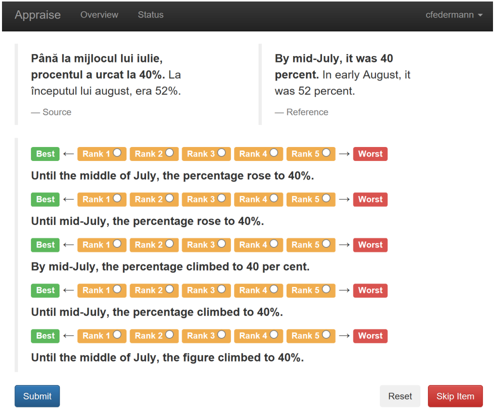
Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Human evaluation: Ranking

Given *N* translation output and source, order them from best to worst.



Human evaluation: Agreement

Evaluators disagree in their assessments.

Inter-evaluator agreement can be measured with Kappa coefficient:

$$K = \frac{p(A) - p(E)}{1 - p(E)}$$

$p(A)$ = proportion of agreement

$p(E)$ = agreement by chance

Ranking provides more consistent results than direct assessment.

Evaluating translator efficiency gain

How does the average translation time per sentence change?

- From scratch
- Using only translation memory
- Between different MT systems

Challenges:

- Translators have different experience and ways of working
- High variability between translation segments
- Easiest cases often solved by translation memories
- How to present the translation in the UI

Needs lots of data or complicated setup and advanced analysis (e.g. mixed-effect regression models).

Why automatic evaluation?

Manual evaluation is expensive

MT researchers rarely have the resources.

Annual competitions (WMT shared tasks) help somewhat.

Manual evaluation is slow

Cannot be used during development.

Especially not for optimization of model parameters and hyperparameters.

Challenges in automatic evaluation

Why is MT evaluation more difficult than ASR evaluation?

Why can we not use word error rate (WER)?

Challenges in automatic evaluation

Multiple correct answers: Ideally there should be several reference translations made by different persons.

Graded correctness: Word choices, grammatical correctness, emphasis (“koira jahtasi kissaa” vs. “kissaa koira jahtasi”), style (“kick the bucket” vs. “die”), ...

Usefulness depends on intended use.

- Translator’s tool: Long segments that require no changes
- Skimming: Meaning should be correct; fluent enough for easy understanding
- Information retrieval: Terminology important; fluency and grammatical correctness do not matter

Global edit distance metrics

Word and letter error rates do not account for possible variations in word order.

Edit distance with moves is an NP-hard problem.

Solutions:

- TER: Shift operation + greedy search (Snover et al. 2006)
- SPEDE: Limited-distance word swapping (Wang and Manning 2012)

Local metrics

Concentrate on small parts of the full text at a time.

Similarity to IR metrics:

- **Precision:** Every item should be found in the reference.
- **Recall:** Anything in the reference should not be left out.

Observing individual words is not adequate (word order!)

Local metrics: BLEU

BLEU ("Bilingual Evaluation Understudy") (Papineni et al. 2002) was one of the first metrics to report high correlation with human judgments of quality.

$$\text{BLEU} = \min\left(1, \frac{\texttt{output-length}}{\texttt{reference-length}}\right)\left(\prod_{i=1}^{4} \texttt{precision}_i\right)^{\frac{1}{4}}$$

Typically calculated over entire corpus (system-level evaluation).

Example in the excercise session.

Local metrics: Problems in BLEU

Does not work for languages with no word boundaries.

Single word or n-gram is scored 0 or 1.

- Inflections: "translation" vs. "translations"
- Derivations: "[he] made translations" vs. "[he] translated"
- Compounds: "Arbeits Geberverband" vs. "Arbeitgeberverband" *(employers' organization)*

Poor measure of adequacy for morphologically rich languages.

Beyond word-based metrics

Preprocessing (stemming, morphological segmentation)

- METEOR (Banerjee and Lavie 2005; Denkowski and Lavie 2011)
- AMBER (Chen and Kuhn 2011)

Characted-based measures

- char-BLEU (Denoual and Lepage 2005)
- Weighted character F-score (chrF3) (Popović 2015)

Combine with word similarity calculation

- Alignment based on character similarity (Homola, Kuboň, and Pecina 2009)
- Tolerant BLEU (Libovický and Pecina 2014)
- LeBLEU (Virpioja and Grönroos 2015)

Semantic similarity using contextual embeddings

- BERTscore (Zhang et al. 2019)
- COMET (Rei et al. 2020)

How to evaluate evaluation metrics?

Goals

- Correct: better systems have higher scores
- Interpretable: intuitive interpretation of translation quality
- Consistent: repeated use gives the same results
- Low cost: efficient computation, no extra work or linguistic resources needed
- Tuning compatible: can be used to tune translation systems

WMT Metrics shared task

- Long-running comparison of evaluation metrics
- Correlation with human evaluation scores

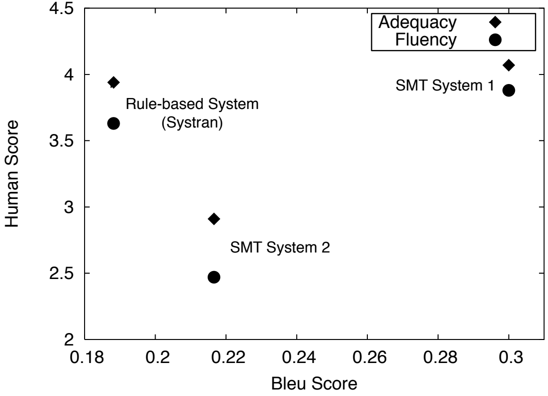
How to evaluate evaluation metrics?

Even if a metric works for comparing similar MT systems, it should not to be trusted for comparing very different ones.

Example from `http://www.statmt.org/book/`:

Evidence of Shortcomings of Automatic Metrics

Rule-based vs. statistical systems



NMT quality on par with human translators?

Sometimes human evaluation has indicated that NMT would be on the level of human translation.

E.g. paper by Microsoft Research:
"Achieving Human Parity on Automatic Chinese to English News Translation" (Hassan Awadalla et al. 2018)

- Direct assessment (score 0-100) by bilingual humans.
- No statistically significant difference between NMT output and reference translations by humans!

NMT quality on par with human translators?

Caveats:

- Are the human translators professionals? Are they translating to their native language?
- How about the human evaluators?
  - Do they understand what to judge (e.g. fluency vs. adequacy)? Even bad NMT is fluent.
  - Skill and time spent: ability to notice subtle differences.
  - Bilingual vs evaluators only speaking target language (use source, or only reference?)
  - Is the document context available?

See e.g. https://www.linkedin.com/pulse/microsoft-mt-reaches-parity-bad-human-translation-tommi-nieminen

or (Toral et al. 2018; Läubli et al. 2020)

Bibliography

Snover, Matthew et al. (2006). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231. URL: http://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf.

Wang, Mengqiu and Christopher Manning (2012). "SPEDE: Probabilistic Edit Distance Metrics for MT Evaluation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, pp. 76–83. URL: http://www.aclweb.org/anthology/W12-3107.

Papineni, Kishore et al. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, pp. 311–318. URL: http://www.aclweb.org/anthology/P02-1040.

Bibliography (cont.)

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: http://www.aclweb.org/anthology/W/W05/W05-0909.

Denkowski, Michael and Alon Lavie (2011). "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 85–91. URL: http://www.aclweb.org/anthology/W11-2107.

Chen, Boxing and Roland Kuhn (2011). "AMBER: A Modified BLEU, Enhanced Ranking Metric". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 71–77. URL: http://www.aclweb.org/anthology/W11-2105.

Bibliography (cont.)

Denoual, Etienne and Yves Lepage (2005). “BLEU in characters: towards automatic MT evaluation in languages without word delimiters”. In: *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing*, pp. 81–86. URL: https://www.aclweb.org/anthology/I/I05/I05-2014.pdf.

Popović, Maja (2015). “chrF: character n-gram F-score for automatic MT evaluation”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. URL: http://aclweb.org/anthology/W15-3049.

Homola, Petr, Vladislav Kuboň, and Pavel Pecina (2009). “A Simple Automatic MT Evaluation Metric”. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics, pp. 33–36. URL: http://www.aclweb.org/anthology/W09-0403.

Bibliography (cont.)

Libovický, Jindřich and Pavel Pecina (2014). “Tolerant BLEU: a Submission to the WMT14 Metrics Task”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 409–413. URL: http://www.aclweb.org/anthology/W14-3353.

Virpioja, Sami and Stig-Arne Grönroos (2015). “LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 411–416. URL: http://aclweb.org/anthology/W15-3052.

Zhang, Tianyi et al. (2019). “BERTScore: Evaluating Text Generation with BERT”. In: *International Conference on Learning Representations*.

Bibliography (cont.)

Rei, Ricardo et al. (Nov. 2020). "COMET: A Neural Framework for MT Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: https://aclanthology.org/2020.emnlp-main.213.

Hassan Awadalla, Hany et al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: *ArXiv e-prints*. URL: https://arxiv.org/abs/1803.05567.

Toral, Antonio et al. (2018). "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation (WMT): Research Papers*, pp. 113–123. URL: https://www.aclweb.org/anthology/W18-6312.

Läubli, Samuel et al. (2020). "A Set of Recommendations for Assessing Human–Machine Parity in Language Translation". In: *Journal of Artificial Intelligence Research* 67, pp. 653–672.

How good is ChatGPT at translation?

System	Zh⇒En	En⇒Zh	De⇒Zh	Ro⇒Zh
Google	31.66	43.58	38.71	39.05
DeepL	31.22	44.31	40.46	38.95
Tencent	29.69	46.06	40.66	n/a
ChatGPT (Direct)	24.73	38.27	34.46	30.84
ChatGPT ($\text{Direct}_{\text{new}}$)	n/a	n/a	30.76	27.51
ChatGPT ($\text{Pivot}_{\text{new}}$)	n/a	n/a	34.68	34.19
GPT-4	28.50	42.50	38.16	37.84

Figure 0: Translation performance of GPT-4 (Date: 2023.03.15).

(Jiao et al. 2023)