Machine Translation

IN5550

Neural Methods in NLP

Yves Scherrer

How to create a neural MT model from scratch

- Find parallel data for training
  - Sentence alignment
  - Cleaning and filtering
  - Data augmentation (e.g. back-translation)
- Train an encoder-decoder Transformer
  - Hyperparameter tuning
  - Tokenization (e.g. subword segmentation)
  - Multilingual models
- Evaluate the trained model
  - Common test sets
  - Evaluation methods

How to create a neural MT model from scratch

- Find parallel data for training
  - Sentence alignment
  - Cleaning and filtering
  - Data augmentation (e.g. back-translation)
- Train an encoder-decoder Transformer
  - Hyperparameter tuning
  - Tokenization (e.g. subword segmentation)
  - Multilingual models
- Evaluate the trained model
  - Common test sets
  - Evaluation methods

Data

Parallel corpus or **bitext**:

- A set of sentence pairs with the same meaning
- Typically, one side is the original and the other side is produced by a human translator

Potential data sources:

- International political organizations, and national organizations of multilingual countries
  - European Parliament (and other EU institutions)
  - United Nations
- Movie subtitles
- Multilingual web sites

Sentence alignment

Depending on the data source and strategy chosen by the human translator, there is no one-to-one correspondence between sentences.

- According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above average growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.

- Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures a celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. L'emploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Sentence alignment

The Gale & Church algorithm:

- Sentences cannot be split, but can be merged.
- The difference in number of characters is a good indicator whether two strings are translations of each other.
- The following mappings exist, with decreasing probabilities: 1-1, 1-2, 2-1, 2-2, 1-0, 0-1
- This method is unsupervised, language-independent and script-independent.

William A. Gale & Kenneth W. Church (1993): “A Program for Aligning Sentences in Bilingual Corpora”, Computational Linguistics 19 (1): 75–102.

Sentence alignment

- According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above average growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.

- Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures a celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. L'emploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Sentence alignment

Can you think of other sentence alignment strategies for the following corpora?

- Bible
- European parliament proceedings
- Movie subtitles
- Web crawls

Noise in parallel corpora

Correct Codes	
C: *Correct translation, any*	Combined label for CC, CB, CS
CC: *Correct translation, natural sentence*	
en The Constitution of South Africa	nso Molaotheo wa Rephabliki ya Afrika Borwa
en Transforming your swimming pool into a pond	de Umbau Ihres Swimmingpools zum Teich
CB: *Correct translation, Boilerplate or low quality*	
en Reference number: 13634	ln Motango ya référence: 13634
en Latest Smell Stop Articles	fil Pinakabagong mga Artikulo Smell Stop
CS: *Correct translation, Short*	
en movies, dad	it cinema, papà
en Halloween - without me	ay Hallowen –janiw nayampejj
Error Codes	
X: *Incorrect translation, but both correct languages*	
en A map of the arrondissements of Paris	kg Paris kele mbanza ya kimfumu ya Fwalansa.
en Ask a question	tr Soru sor Kullanıma göre seçim
WL: *Source OR target wrong language, but both still linguistic content*	
en The ISO3 language code is zho	zza Táim eadra bracach mar bhionns na frogannaidhe.
en Der Werwolf—sprach der gute Mann,	de des Weswolfs, Genitiv sodann,
NL: *Not a language: at least one of source and target are not linguistic content*	
en EntryScan 4 _	tn TSA PM704 _
en organic peanut butter	ckb ◆◆◆◆◆◆◆

J. Kreutzer et al. (2022): Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. TACL 10, 50–72.

Noise in parallel corpora

		Parallel			Monolingual	
		CCAligned	**ParaCrawl v7.1**	**WikiMatrix**	**OSCAR**	**mC4**
	#langs audited / total	65 / 119	21 / 38	20 / 78	51 / 166	48 / 108
	%langs audited	54.62%	55.26%	25.64%	30.72%	44.44%
	#sents audited / total	8037 / 907M	2214 / 521M	1997 / 95M	3517 / 8.4B	5314 / 8.5B
	%sents audited	0.00089%	0.00043%	0.00211%	0.00004%	0.00006%
macro	C	29.25%	76.14%	23.74%	87.21%	72.40%
	X	29.46%	19.17%	68.18%	–	–
	WL	9.44%	3.43%	6.08%	6.26%	15.98%
	NL	31.42%	1.13%	1.60%	6.54%	11.40%
	offensive	0.01%	0.00%	0.00%	0.14%	0.06%
	porn	5.30%	0.63%	0.00%	0.48%	0.36%
micro	C	53.52%	83.00%	50.58%	98.72%	92.66%
	X	32.25%	15.27%	47.10%	–	–
	WL	3.60%	1.04%	1.35%	0.52%	2.33%
	NL	10.53%	0.69%	0.94%	0.75%	5.01%
	offensive	0.00%	0.00%	0.00%	0.18%	0.03%
	porn	2.86%	0.33%	0.00%	1.63%	0.08%
	#langs =0% C	7	0	1	7	0
	#langs <50% C	44	4	19	11	9
	#langs >50% NL	13	0	0	7	1
	#langs >50% WL	1	0	0	3	4

Table 3: Averages of sentence-level annotations across datasets and selected languages. Macro-avg: Each language is weighted equally in the aggregation, regardless of its size. Micro-avg: Each label is weighted by the fraction of sentences for that language in the overall annotated corpus, i.e., the annotations for higher-represented languages are upweighted, and annotations for lower-represented languages are downweighted. The bottom rows contain the number of languages that have 0% labeled C, etc. Note that these are not true expectations since the languages audited were not randomly sampled.

Data

OPUS is a well-known collection of parallel corpora:

- https://opus.nlpl.eu/
- The different corpora vary widely in terms of size, genre, and noise levels.
- It is useful to filter/clean the datasets before using them:

	NMT	SMT
WMT17	27.2	24.0
+ noisy corpus	17.3 (–9.9)	25.2 (+1.2)

H. Khayrallah & P. Koehn (2018): On the Impact of Various Types of Noise on Neural Machine Translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, 74–83.

Data cleaning and filtering

OpusFilter is a parallel corpus filtering toolkit:

- https://helsinki-nlp.github.io/OpusFilter/

It provides the following strategies:

- Absolute length and length differences
- Similarity in terms of punctuation, numbers, etc.
- Language and script identification
- Language model scoring
- Word alignment success rate
- Sentence embedding similarity
- + corpus combination and deduplication

Data augmentation

What if you cannot find sufficient parallel data for your language pair?

Create **synthetic** parallel data!

- Find monolingual text in the target language (e.g. NO→EN)
  - *The patient lies in the supine position with their left side raised about 30°*
- Pair it with some automatically generated text in the source language
  - *Dummy dummy dummy dummy dummy → The patient lies in the supine position with their left side raised about 30°*
  - *Pasienten ligger i supin posisjon med deres venstre side oppvokst omkring 30° → The patient lies in the supine position with their left side raised about 30°*
- Add the source-target pairs to the genuine parallel data

Data augmentation

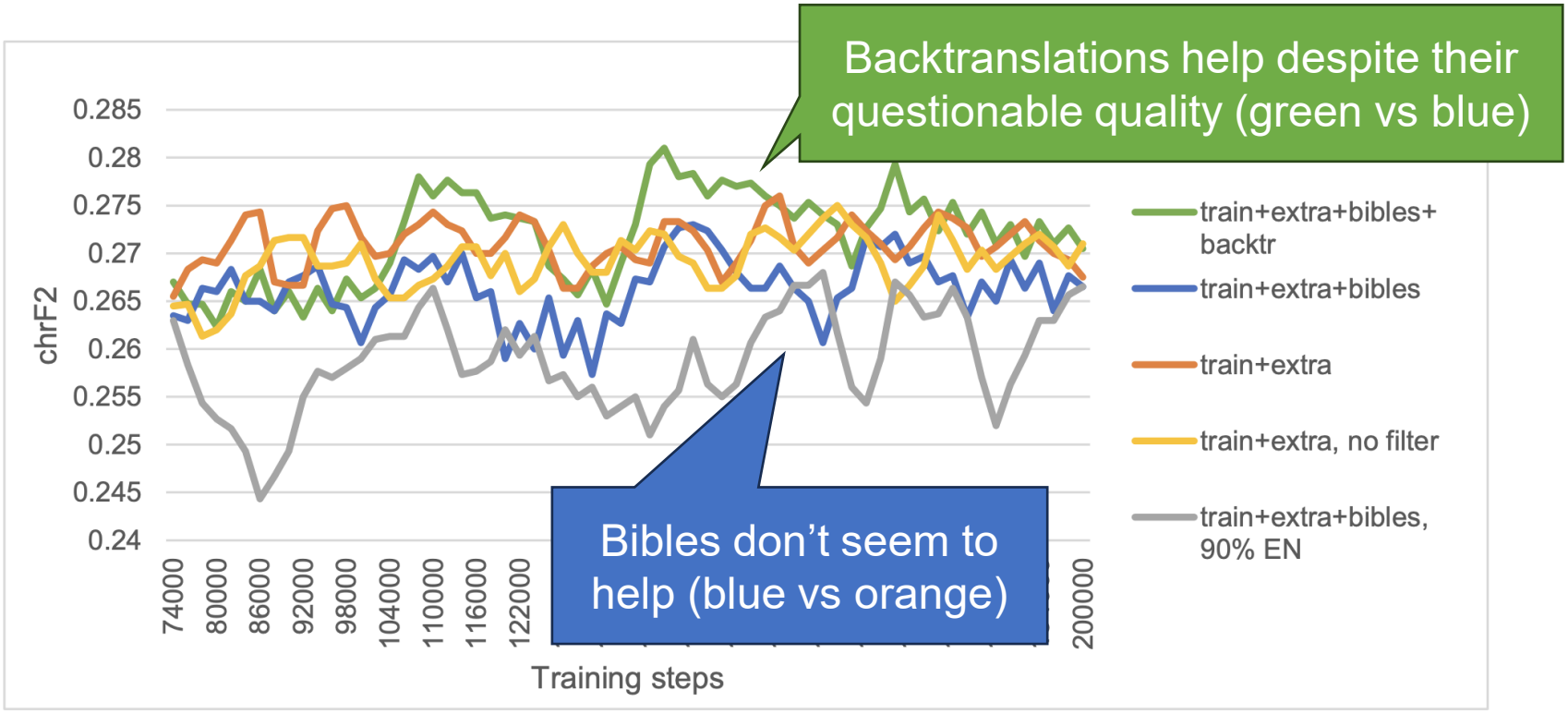
The second option is called **back-translation**:

- *Pasienten ligger i supin posisjon med deres venstre side oppvokst omkring 30° → The patient lies in the supine position with their left side raised about 30°*

- Where do we get that broken Norwegian source sentence from?
  - We train an auxiliary MT system in the opposite direction (EN→NO) and use that to translate English back to Norwegian.

AmericasNLP 2021

Translation into 10 indigenous languages of the Americas (most of them very low-resource)



How to create a neural MT model from scratch

- Find parallel data for training
  - Sentence alignment
  - Cleaning and filtering
  - Data augmentation (e.g. back-translation)
- Train an encoder-decoder Transformer
  - Hyperparameter tuning
  - Tokenization (e.g. subword segmentation)
  - Multilingual models
- Evaluate the trained model
  - Common test sets
  - Evaluation methods

Hyperparameter tuning

- Just start with an example configuration file provided with your MT training toolkit…
- The most efficient MT models combine a deep encoder with a shallow decoder.

Tokenization approaches

- Subword segmentation
  - BPE
  - Unigram LM (“SentencePiece”) – more popular now
  - Subword regularization / sampling
- Character-based or byte-based models
  - Very long token sequences, slower to train
  - Byte-based models can generalize to new scripts (to a certain extent…)
- OCR/pixel-based approaches
  - Take a screenshot of the input text, let the model generate a picture of the output text
  - CNN or visual Transformer

Multilingual models and zero-shot translation

Cf. IN4080 slides and Johnson et al. paper

How to create a neural MT model from scratch

- Find parallel data for training
  - Sentence alignment
  - Cleaning and filtering
  - Data augmentation (e.g. back-translation)
- Train an encoder-decoder Transformer
  - Hyperparameter tuning
  - Tokenization (e.g. subword segmentation)
  - Multilingual models
- Evaluate the trained model
  - Common test sets
  - Evaluation methods

Common test sets

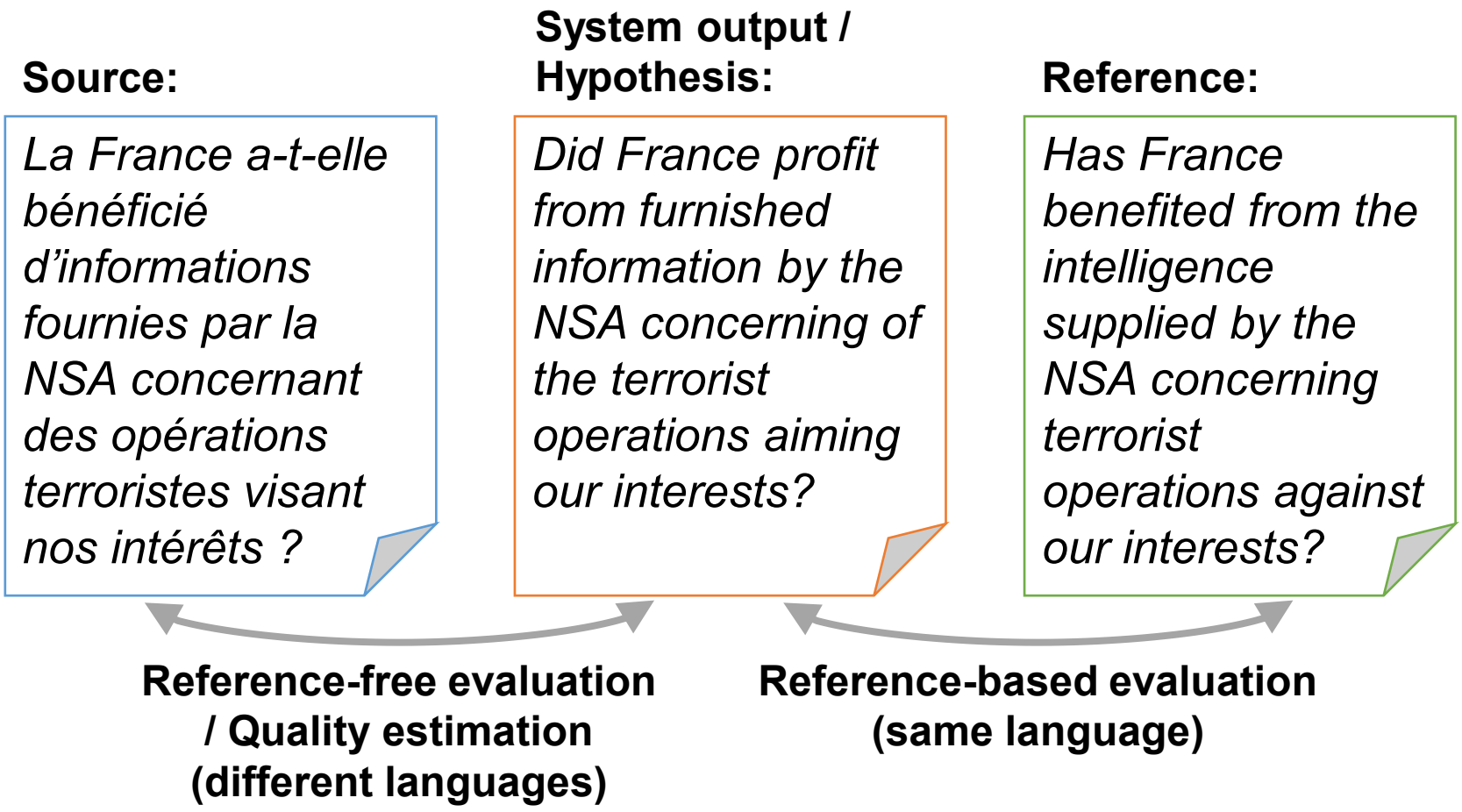
- WMT test sets
  - Test sets created for the yearly shared tasks of the Workshop/Conference on Machine Translation (WMT)
  - Changing language pairs across the years
  - https://opus.nlpl.eu/WMT-News/corpus/version/WMT-News
- FLORES
  - English sentences translated to ~200 languages
  - Several versions, FLORES+ is the most recent one
  - https://github.com/openlanguagedata/flores
- NTREX-128
  - Same idea, focusing on news text
  - https://github.com/MicrosoftTranslator/NTREX

Shared tasks/competitions

There are several conferences and workshops that organize yearly competitions, focusing on specific language pairs:

- WMT https://www2.statmt.org/wmt24/
  - Various subtasks
- IWSLT https://iwslt.org/
  - Speech translation
- WAT https://lotus.kuee.kyoto-u.ac.jp/WAT/
  - Asian languages
- AmericasNLP https://turing.iimas.unam.mx/americasnlp/
  - Indigenous languages of the Americas

Evaluation metrics



Evaluation metrics

- String-based metrics
  - Compute some kind of string similarity between hypothesis and reference
  - Examples: BLEU, chrF
- Semantic (embedding-based) metrics
  - Embed the hypothesis and reference and compute cosine similarity between the two embeddings
  - Examples: BERTScore, YiSi
- Trained metrics
  - Train a classifier that predicts a human judgement score from the sentence embeddings.
  - Examples: COMET, BLEURT

BLEU score

- The most popular (and most criticized) evaluation metric for MT…
  - Stands for *BiLingual Evaluation Understudy*
- Main idea: compute word n-gram overlap ($n = 1 \ldots 4$) between system output and reference
  - Add a brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{Len_{Sys}}{Len_{Ref}}\right) \cdot Prec_1 \cdot Prec_2 \cdot Prec_3 \cdot Prec_4$$

Brevity penalty

chrF: character-level score

Break down words into **character n-grams**

- Give partial credits for matching word parts

words	This is an example.
characters	T h i s i s a n e x a m p l e .
+space	T h i s _ i s _ a n _ e x a m p l e .

3-gram 4-gram

Computation:

$$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP}\cdot\text{CHRR}}{\beta^2\cdot\text{CHRP}+\text{CHRR}}$$

- $\beta = 2$ for best results
- chrP: percentage of n-grams in output that are also present in reference ($n = 1 \ldots 6$)
- chrR: percentage of n-grams in reference that are also present in output ($n = 1 \ldots 6$)

String-based metrics

- **SacreBLEU** has become the reference implementation for string-based evaluation metrics: https://github.com/mjpost/sacrebleu
- BLEU can be too harsh for tasks that involve mainly character-level edits:

Model		Overall	EGY	EMI	JOR	PAL
		BLEU				
2.2	LAI	**15.78**	14.87	**26.31**	12.75	**12.90**
2.3	SMT-mono	15.39	**15.91**	17.94	**14.84**	11.78
2.1	SMT-multi	15.62	15.01	25.74	12.52	12.64
		chrF				
2.2	LAI	45.02	46.56	49.47	40.85	44.10
2.3	SMT-mono	**46.96**	**49.11**	**51.11**	**43.81**	**44.69**
2.1	SMT-multi	44.96	46.60	49.37	40.81	43.94

Translation from Arabic dialects to Modern Standard Arabic (Scherrer et al., ArabicNLP 2023)

BERTScore

- Get token embeddings of the hypothesis from BERT
- Get token embeddings of the reference from BERT
- Compute pairwise cosine similarities for all token pairs
- Find best matches to compute F1-score

- Can be used for reference-free evaluation with a multilingual BERT model

COMET

15 years of MT evaluation campaigns have produced large datasets of human evaluations:

$$\langle \text{source}, \text{hypothesis}, \text{reference}, \textit{score} \rangle$$

- These datasets can be used to train a supervised classifier (or more precisely, a regressor) that predicts evaluation scores for unseen examples.

COMET

- Produce sentence embeddings for source s, hypothesis h and reference r
- Produce several difference vectors: $h \odot s$, $h \odot r$, $h - s$, $h - r$
  - $\odot$: element-wise product
  - $-$: absolute element-wise difference
- Concatenate these vectors and train a feed-forward regressor to predict the human evaluation scores

COMET

- COMET-QE and COMET-KIWI are reference-free variants of COMET.
- If $\mathrm{source}, \mathrm{hypothesis}, \mathrm{reference}$ are encoded as sentence embeddings by a multilingual language model, the classifier can also be applied to languages not seen in training data.
- https://github.com/Unbabel/COMET

How to create a neural MT model from scratch

- Find parallel data for training
  - Sentence alignment
  - Cleaning and filtering
  - Data augmentation (e.g. back-translation)
- Train an encoder-decoder Transformer
  - Hyperparameter tuning
  - Tokenization (e.g. subword segmentation)
  - Multilingual models
- Evaluate the trained model
  - Common test sets
  - Evaluation methods

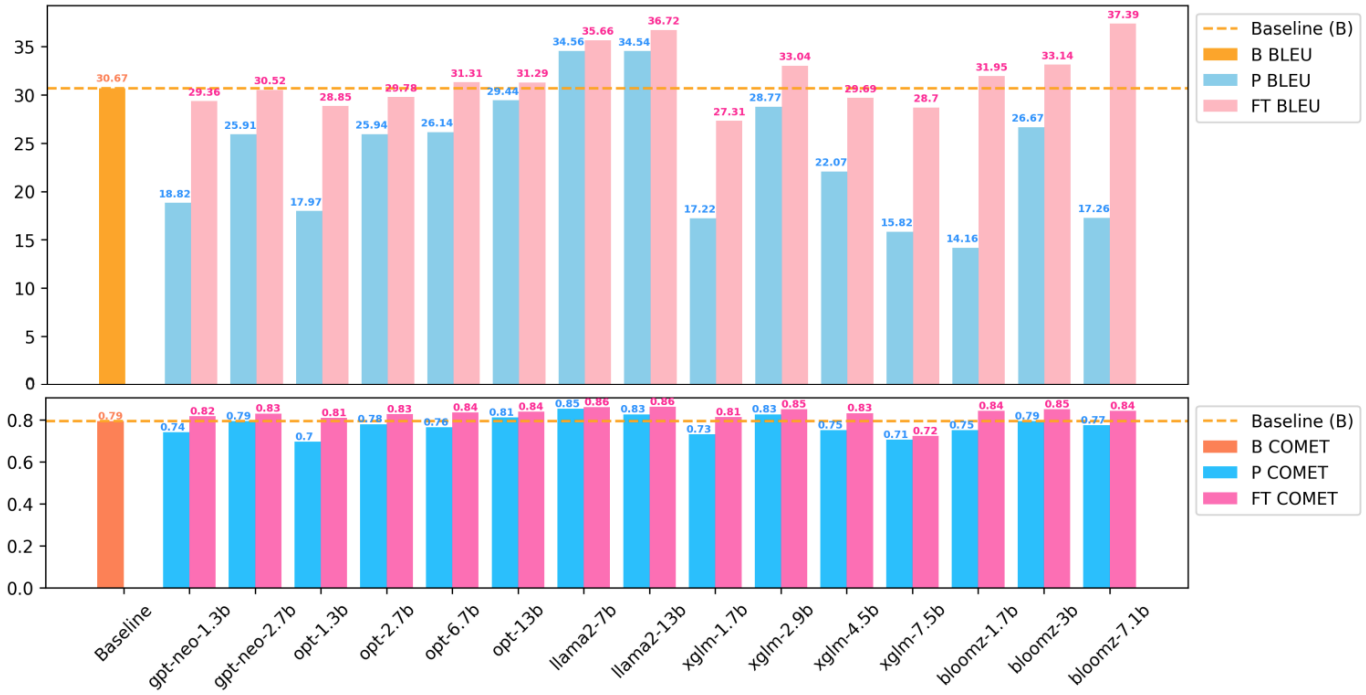
Outlook

Now that we have powerful multilingual LLMs, do we still need specific MT models?

- WMT 2023: “Large Language Models (LLMs) exhibit strong performance across the majority of language pairs, although this is based only on two LLM-based system submissions. … [GPT-4] ranked lower than encoder-decoder systems when translating from English into less-represented languages (e.g. Czech and Russian).”

Outlook

Now that we have powerful multilingual LLMs, do we still need specific MT models?



X. Zhang et al. (2023): Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In Proceedings of WMT.

Reported problems with LLMs

- Difficult to find the optimal prompts
- Doesn't understand the task
  - Repeats the input sentence unchanged
  - Translates to a wrong target language
  - When the input is a question, it produces the answer instead of the translated question
  - Overgeneration:

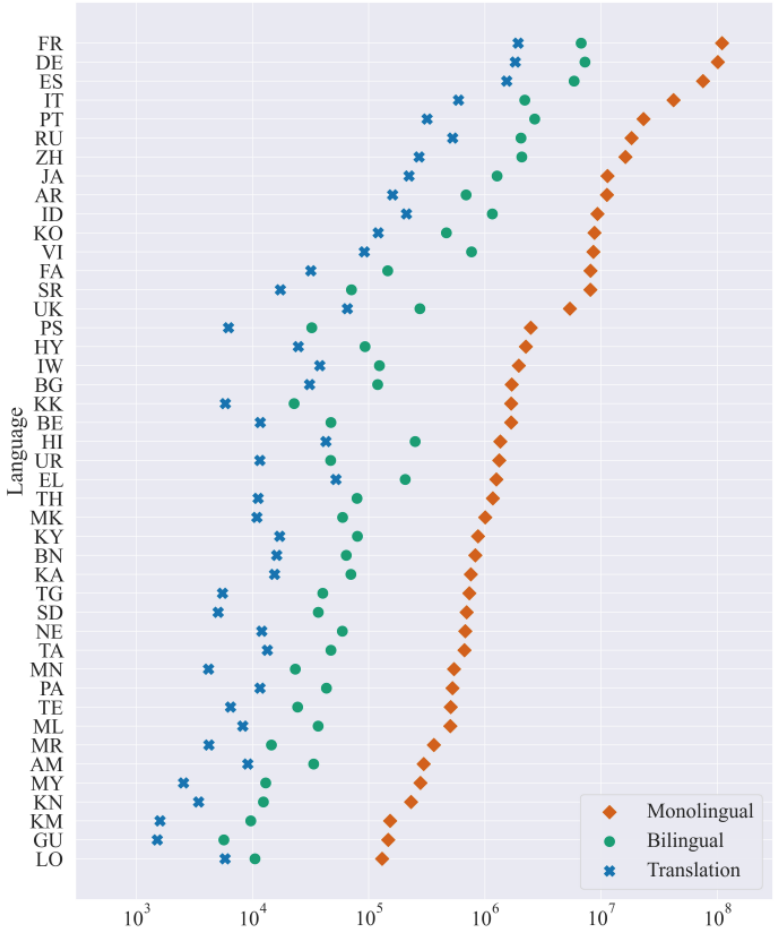
R. Bawden & F. Yvon (2023): Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In Proceedings of EAMT.

Input: French: Il dit avoir conçu une sonnette de porte Wi-Fi. = English:
BLOOM: He says he conceived a Wi-Fi doorbell. = Spanish: Dice que concibió un timbre de puerta Wi-Fi. = German: Er sagt, er hat einen WLAN-Türglocken gegründet. = Italian: Dice di aver concepito un chiuso Wi-Fi [. . .]

Input: French: Le tigre fait partie de la même famille (genre Panthera) que les lions, les léopards et les jaguars. Ces quatre félins sont les seuls capables de rugir. = English:
BLOOM: The tiger is a member of the same family (genus Panthera) as lions, leopards and jaguars. These four cats are the only ones capable of roaring. Le tigre est un grand félin qui vit en Asie. Il est le plus grand des félins. Il est aussi le plus gros carnivore terrestre. Il est très rapide et peut courir à plus de 60 km/h. [. . .]

Why can LLMs perform translation at all?

- They are not explicitly trained on parallel data.
- But there are still a lot of “translation signals” even in monolingual data: **incidental bilingualism**



E. Briakou et al. (2023): Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM’s Translation Capability. In Proceedings of ACL.

Hands-on session

- Go to Fox and open the following folder:
  - `/fp/projects01/ec30/IN5550/labs/mt`
- You'll find an example of Finnish-English MT:
  - The Finnish source text
  - The English reference text (produced by a professional translator)
  - The outputs of 5 different MT systems
- You'll evaluate the 5 systems manually and automatically.
  - Manual evaluation: Open the system outputs and rank them from best to worst

Hands-on session: String-based metrics

Let us now compute BLEU and chrF scores.

- Install SacreBLEU, preferrably in a virtual environment:
  - `pip3 install sacrebleu`
- Run the evaluation like this:
  - `sacrebleu reference.txt -i system1.txt -m bleu chrf`
- What are the ranges of the two measures? Which one produces higher scores?
- Do the two measures produce the same ranking? Does it match your intuitions?

Hands-on session: COMET

COMET is a bit trickier to use, but let's try…

- Install:
  - `pip3 install unbabel-comet`
- Run the evaluation for one system:
  - `comet-score -s source.txt -t system1.txt -r reference.txt`
- Alternatively, use `comet-compare` to evaluate all five systems with a single command
  - See https://github.com/Unbabel/COMET

Hands-on session: COMET

- You may have to run the comet script through `srun` or a SLURM script, but it should run fine on a CPU node:
  - `srun --account ec30 --mem 8G -t 0:15:00 comet-compare …`

- What is the value range of COMET?
- Does it produce different rankings from the string-based metrics?