When I look at an article in Russian, I say: This is really written in English, but has been coded in some strange symbols; I will now proceed to decode.

Warren Weaver, Excerpt from a letter written in 1955

UNSW SYDNEY

1

Natural Language Processing (NLP)

COMP6713 – 2025 Term 1

Convener

Dr. Aditya Joshi

aditya.joshi@unsw.edu.au

Week 8

Machine Translation

Schedule

2025 Term 1

UNSW SYDNEY

2

UNSW SYDNEY | Australia's Global University

Week 8
Machine Translation

Introduction
Task definition
Terminology

MT Evaluation
BLEU
BLEURT
Other metrics

SMT
Alignment
Language model
IBM Models
Moses

NMT
Transformer decoding
Pivot-based MT
Subword-based MT
Post-editing
....

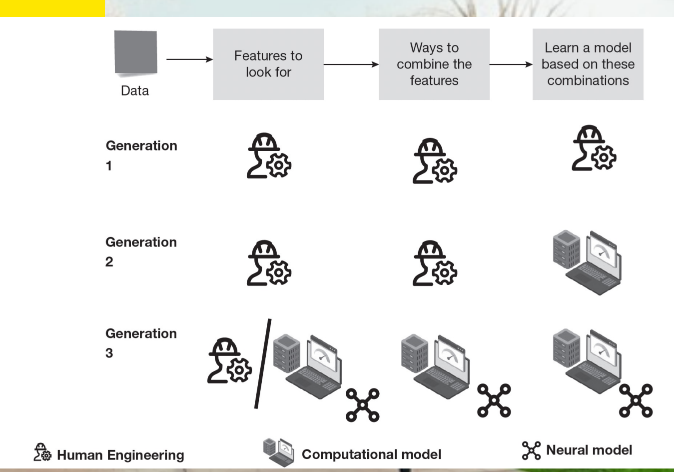
Chapter 7 of Bhattacharyya, Joshi, 'Natural Language Processing', Wiley, 2023.

Chapter 13 of Jurafsky, Martin, 'Speech and Language Processing', https://web.stanford.edu/~jurafsky/slp3/13.pdf

3

UNSW SYDNEY | Australia's Global University

Topics this week



Statistical Machine Translation
Evaluation Metrics

Transformer decoding
NMT
LLM-based translation

4

UNSW SYDNEY
Australia's Global University

Part 0

Introduction

5

Good translation is helpful

https://www.pcmag.com/news/google-lens-now-translates-offline

6

Machine Translation (MT)

Objective: Translate text from a **source** language to a **target** language such that the meaning is preserved.

English *The boy drinks milk.* → MT → *Il ragazzo beve il latte.* Italian

1 2 3 4 → 1 2 3 X 4

English *The boy drinks milk.* → MT → *Ólann an buachaill bainne.* Irish

1 2 3 4 → 3 1 2 4

English *The boy drinks milk.* → MT → *mulga doodh pito* Marathi

1 2 3 4 → 2 4 3

Represents sequence-to-sequence tasks in NLP

Motivation for Transformer (weeks 3 and 4)

UNSW SYDNEY

7

Requirements of MT

Learn to translate words.

English *The boy drinks milk.* → MT → *Il ragazzo beve il latte.* Italian

1 2 3 4 → 1 2 3 ~~X~~ 4

One source language word may generate more than (or even no) target language word.

English *The boy drinks milk.* → MT → *Ólann an buachaill bainne.* Irish

1 2 3 4 → 3 1 2 4

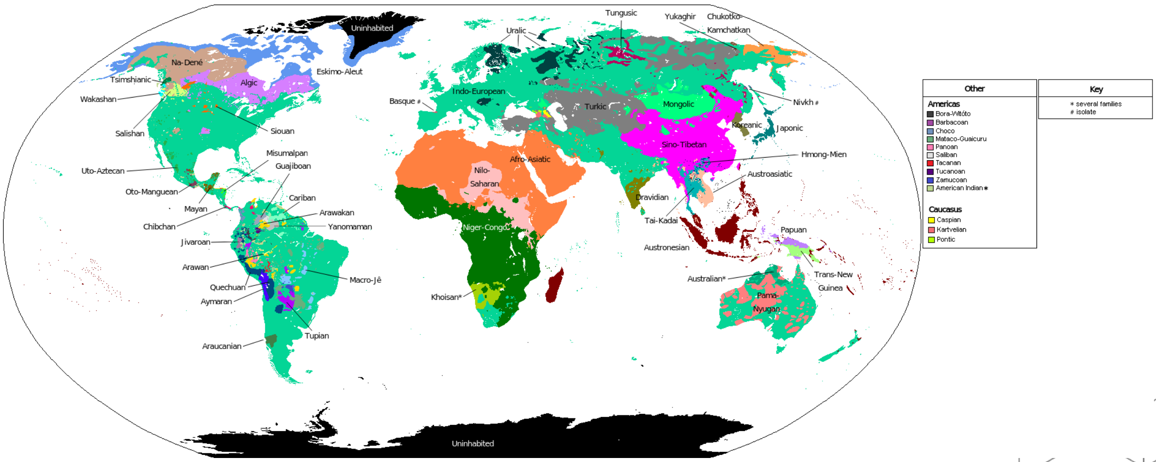
Place them in the correct order in the target language.

Some language pairs will require less reordering than others.

UNSW SYDNEY

8

Language families



https://en.wikipedia.org/wiki/Language_family

9

Challenges

Promotional divergence

E: John usually goes home ← S: Juan suele ira casa ('John tends to go home')

Lexical divergence

"sweater", "jumper", "jacket"…

70 words for snow in Eskimo language

Pragmatics

"Tiramisu" -> "pick me up"

Can you think of similar examples in the languages other than English that you speak?

https://en.wikipedia.org/wiki/Eskimo_words_for_snow

10

Exercise

Taking inspiration from fictional languages in pop culture….

English	Language X
The boy eats rice.	ghada mada vadgha
The girl eats rice.	ghadi mada vadghi
The girl drinks water.	bheedi mali vadghi
The boy drinks milk.	bheedi kook vadgha

Parallel corpus: Pairs of sentences in the source and target languages

You learned to form correspondences between words in the two languages. -> Alignment

What is the translation of 'boy' in language X?

You learned the word order in the target language -> Language model

https://ozclo.org.au/wp-content/uploads/2014/11/2008-first-round-problems.pdf
https://movieweb.com/iconic-fictional-lanuages-movies-tv-shows/

11

Why machine translation?

Early MT system: SYSTRAN

Russian to English Translation

Funded by US Air Force

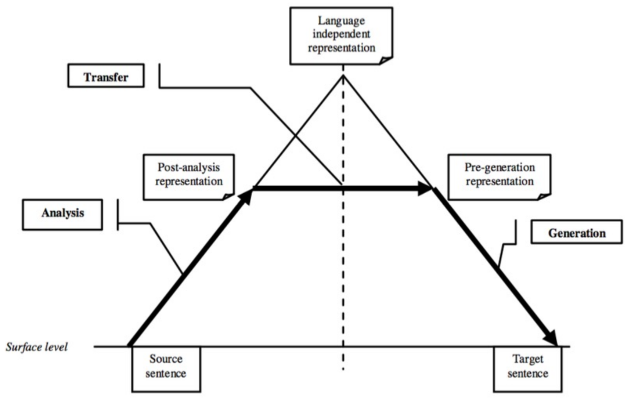
Systems today:

Google, Bing….

Bridge knowledge gap

Make information accessible

"Connect people!"



Rule-based MT: SYSTRAN

12

MT using zero-shot prompting

```
In [1]: from transformers import T5Tokenizer, T5ForConditionalGeneration

        tokenizer = T5Tokenizer.from_pretrained("google/flan-t5-large")
        model = T5ForConditionalGeneration.from_pretrained("google/flan-t5-large")

        input_text = "translate English to German: How old are you?"
        input_ids = tokenizer(input_text, return_tensors="pt").input_ids

        outputs = model.generate(input_ids)
        print(tokenizer.decode(outputs[0]))
```

```
<pad> Wie alte sind Sie?</s>
```

13

Australia's Global University

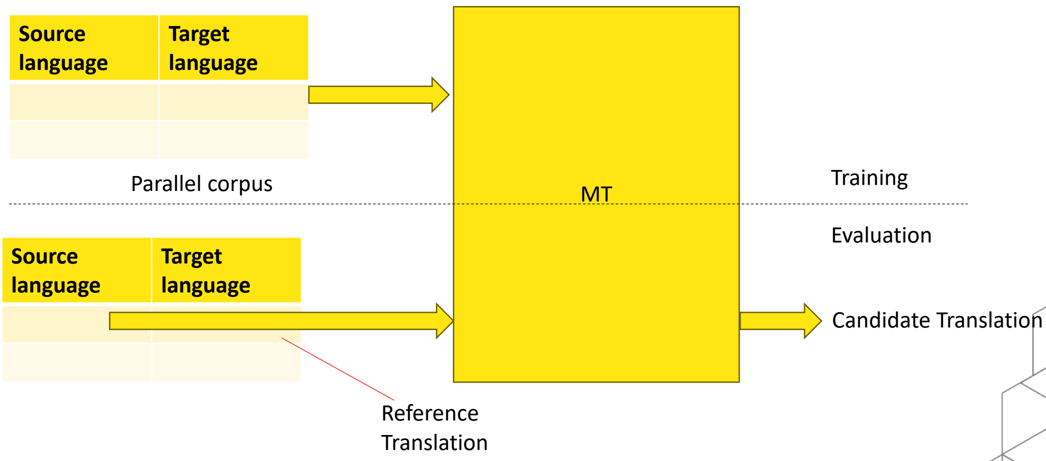
Part 1

MT Evaluation

BLEU: Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 2002.
Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 2020.

14

Evaluating MT systems



15

BLEU

Bilingual Evaluation Understudy

Automatic metric for the evaluation of MT

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BP: Recall expressed as brevity penalty
pn: Clipped n-gram precision
wn: Fixed weight associated with a value of n

Several limitations are known; often used in conjunction with other metrics (coming up soon)

16

We need a mechanism to compare pairs of sentences

Given a reference translation and an output translation, we need a metric that measures the quality of the translation.

Source sentence: The boy drinks milk.
Reference translation: Il ragazzo beve il latte

What are the possible options?

(A) What proportion of words in the output translation occur in the reference translation? (Precision)

(B) What proportion of words in the output translation occur in the reference translation, assuming a word can appear at most as many times it does in the reference? (clipped precision)

Candidate translation	Precision	Clipped Precision
il ragazzo bevo il latte	4/5	4/5
beve beve beve beve beve	5/5	1/5
ragazzo beve beve il il	5/5	3/5

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Clipped precision ignores repeating words.

17

… is precision enough?

Source sentence: The boy drinks milk.
Reference translation: Il ragazzo beve il latte

(C) What proportion of words in the reference translation occur in the candidate translation? (recall)

(D) Precision takes care of matching words. Can we just look at the proportion of reference and candidate sentence lengths? (r/c)

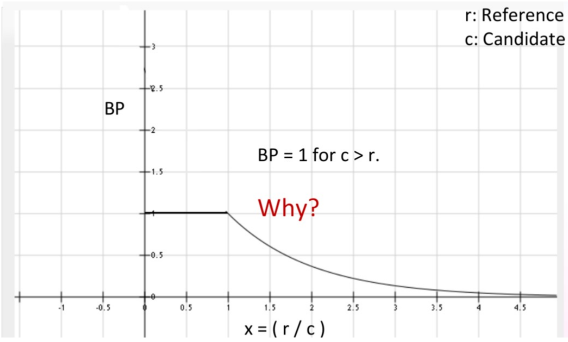
Candidate translation	Precision	Clipped Precision	Recall	r/c
il ragazzo bevo il latte	4/5	4/5	4/5	5/5
il ragazzo	2/2	2/3	2/5	2/5
il ragazzo beve	3/3	3/3	3/5	3/5

18

Recall as a brevity penalty

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

r: Reference sentence length
c: Candidate sentence length



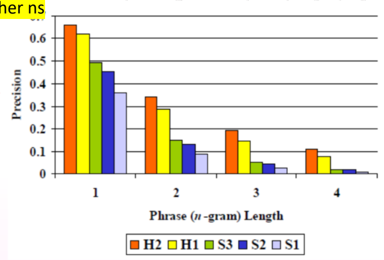
19

The formula for BLEU

$$\mathrm{B{\small LEU}} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Accommodates decay in precision values for higher ns



$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

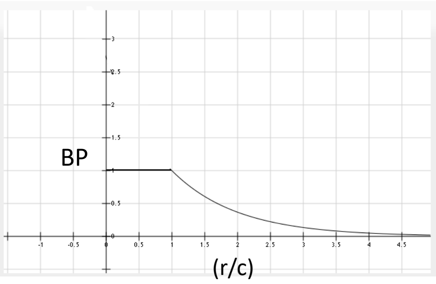
20

Range of BLEU

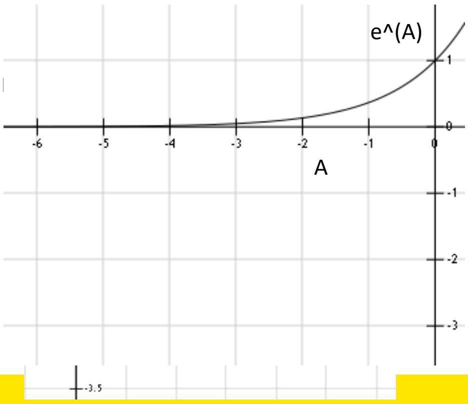
$$\text{BLEU} = \text{BP} \cdot \exp\left(\overset{A}{\boxed{\sum_{n=1}^{N} w_n \log p_n}}\right)$$

p_n : Between 0 and 1
log p_n : Between –(infinity) and 0
A: Between –(infinity) and 0
e ^ (A): Between 0 and 1

BLEU lies between 0 and 1; higher the better.



BP: Between 0 and 1



Graph drawn using www.fooplot.com

21

Other metrics: ROUGE

Recall-Oriented Understudy for Gisting Evaluation

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

ROUGE-N: ROUGE over n-grams

ROUGE-L: Considers longest common subsequence

ROUGE-L: Weighted ROUGE-L: All common subsequences are considered with weight based on length

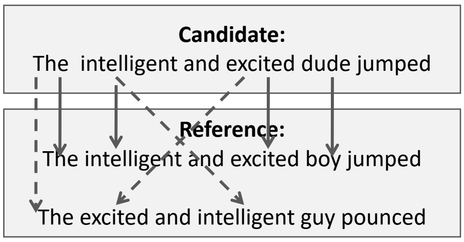
ROUGE-S: Precision/Recall by matching skip bigrams

Chin-yew Lin, "Rouge: a package for automatic evaluation of summaries", 2004

22

Other metrics: METEOR

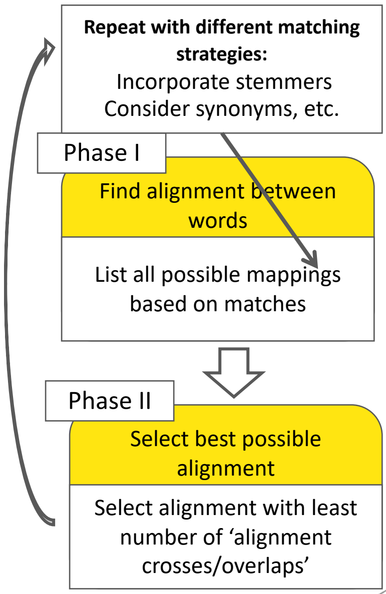
Harmonic mean of precision and recall over unigrams



$$Fmean = \frac{10PR}{R + 9P}$$

$$Score = Fmean * (1 - Penalty)$$

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched}\right)$$



Satanjeev Banerjee and Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005

UNSW SYDNEY

23

Manual Evaluation Strategies

Assigning fluency and adequacy scores on five (Absolute)

Adequacy: *Is the meaning translated correctly?*	***Fluency:*** *Is the sentence grammatically valid?*
5 = Yes	5 = Yes
4 = Mostly	4 = Mostly
3 = Much	3 = Much
2 = Little	2 = Little
1 = None	1 = None

Ranking translated sentences relative to each other (Relative)

Ranking translations of syntactic constituents drawn from the source sentence (Relative)

Automated metrics like BLEU correlate poorly with human judgments. ☹

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder, "(Meta-) Evaluation of Machine Translation", ACL Workshop on Statistical Machine Translation 2007

UNSW SYDNEY

24

BLEU + BERT = BLEURT

Text generation metric based on BERT

Step 1: Create synthetic data for a regression task: sentence pair -> score

sentence1
sentence2
sentence3
....

1. BERT mask-filling
2. Backtranslation
3. Random word-dropping

Perturbation of sentence1
Perturbation of sentence2
Perturbation of sentence3
....

Reference	Candidate	Expected output label
sentence 1	perturb1	BLEU(sentence1, perturb1)
sentence 2	perturb2	BLEU(sentence2, perturb2)
....		

BERT → BERT fine-tuned for regression

UNSW SYDNEY

25

BLEU + BERT = BLEURT

Text generation metric based on BERT

Step 2: Fine-tune on human ratings

Reference	Candidate	Human rating
refsentence 1	candsentence1	hr1
refsentence 2	candsentence2	hr2
....		

BERT → BERT fine-tuned for regression → BERT fine-tuned for regression

Synthetic data based on BLEU scores

Human rating data

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

UNSW SYDNEY

26

In [1]:
```python
from nltk.translate import bleu_score

reference1 = "The cat is out of the bag"
reference2 = "The cat is outside the bag"
```

In [2]:
```python
candidate = "The cat is in the bag"

BLEUscore = bleu_score.sentence_bleu([reference1.split(), reference2.split()], candidate.split())
print(BLEUscore)
```

In [3]:
```python
from rouge_score import rouge_scorer

scorer = rouge_scorer.RougeScorer(['rouge1', 'rougeL'], use_stemmer=True)
scores = scorer.score(reference1, candidate)
```

In [4]:
```python
scores
```

Out[4]:
```
{'rouge1': Score(precision=0.8333333333333334, recall=0.7142857142857143, fmeasure=0.7692307692307692),
 'rougeL': Score(precision=0.8333333333333334, recall=0.7142857142857143, fmeasure=0.7692307692307692)}
```

In [5]:
```python
from bleurt import score
```

In [7]:
```python
scorer = score.BleurtScorer()
scores = scorer.score(references=[reference1], candidates=[candidate])
assert isinstance(scores, list) and len(scores) == 1
print(scores)
```

```
[0.4304013252258301]
```

Available in in-class demos.

27

Part 2

Statistical MT

https://uva-slpl.github.io/nlp2/resources/papers/CollinsIBM.pdf

28

Some parallel datasets

EuroParl

Proceedings of the European Parliament; 21 European languages

United Nations parallel corpus

Six languages: Arabic, Chinese, English, French, Russian, Spanish

OpenSubtitles corpus

Movie and TV subtitles

29

Noisy-channel model

Remember the opening quote of this module.

Noisy channel:

A noisy channel distorts the input signal

Can we recover the input signal based on the output obtained at the end of the noisy channel?

30

31

Language model P(t)

P(t) is the probability of a target language sentence

Joint probability over words

Can be learned from an unsupervised corpus

Reflects fluency of a sentence

Remember: perplexity?

Challenge: Can you think of ways in which P(t) can be used to prefer certain ‘style’ (e.g.: newspaper headlines?)

32

IBM Models

33

Expectation-maximization algorithm

We need alignments to compute the translations… and we need the translations to compute the alignment

Demo time!

Expectation-Maximization (EM) Algorithm

Initialise model parameters

Assign probabilities to missing data : E step

Estimate model parameters from data : M step

Repeat E-M steps

Example EM for MT

Initialise word translation (“tr” function): (a) uniform priors, (b) use a dictionary?

Compute alignments and resultant probabilities

Remember the alignment will be a ‘probability’. Therefore, you will end up needing to update the ‘tr’ function.

Repeat.

34

Part 3

Transformer-based MT

35

Core idea

Many practical NLP tasks can be cast as word prediction
A powerful-enough language model can solve them with a high degree of accuracy

State-of-the-art: Transformer and models derived from it

36

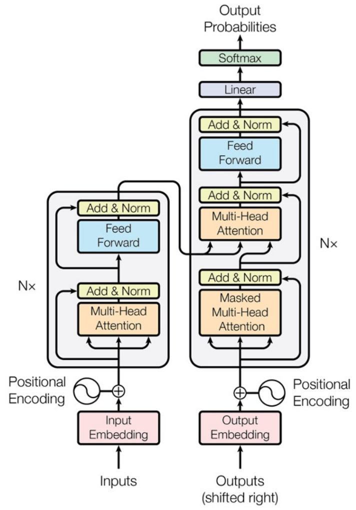
Transformer: Recap

Transformer is a seq2seq architecture (Week 3)

Encoder-Decoder

Multi-head attention

Decoder generates the output sequence



37

Transformer Decoder: Recap

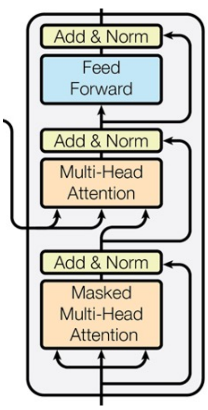
Decoder uses two types of attention:

Masked multi-head attention

For position i, this is computed only over positions 0 to i-1

Multi-head (cross) attention

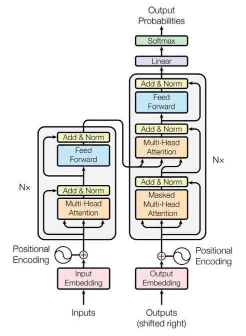
Use output of the encoder to compute attention; apply to the output of the current decoder



38

During training…

Teacher forcing: During training, the *input sequences are always as present* in the training set. The observed output is used to compute loss and update parameters.



the boy drinks milk
il ragazzo beve il latte

Input Sequence	Output Sequence	Expected Output	Observed Output
the boy drinks milk	<s>	il	ragazzo
the boy drinks milk	<s> il	ragazzo	mangia
the boy drinks milk	<s> il ragazzo	beve	il
….	..	….	….

39

Side-note: OpenNMT

Open-source neural machine translation

Maintained by SYSTRAN and Ubiqus

Pytorch: https://github.com/OpenNMT/OpenNMT-py

A good tutorial by Yasmin Moslem: https://github.com/ymoslem/OpenNMT-Tutorial

40

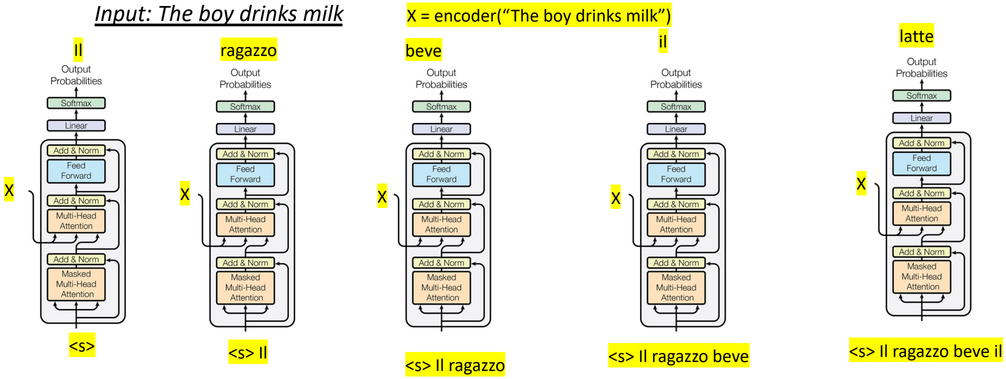
Transformer Decoding (During inference)

Use a pre-trained/fine-tuned Transformer model to generate an output sequence

Auto-regressive decoding: Generate words one after the other (from left to right; or right-to-left)

Input: The boy drinks milk

X = encoder("The boy drinks milk")



41

Decoding

Desirables:

- Variability/diversity of output
- Coherence with respect to input

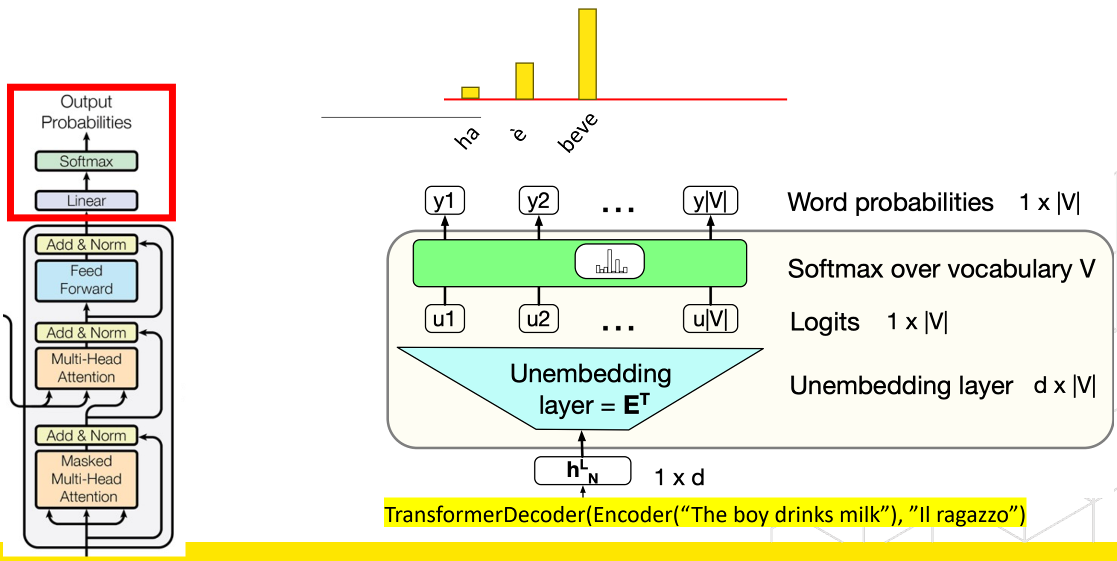
Strategies:

- Greedy decoding
- Decoding by sampling
- Beam search

Refer to Chapter 10 of Jurafsky-Martin for greedy decoding and sampling.

42

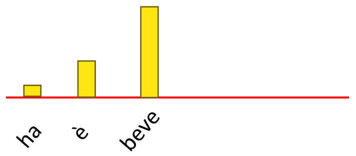
Let's revisit the Language Model Head…



43

Greedy decoding

Generate a word with the highest probability



"beve"

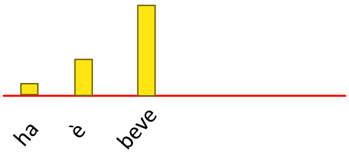
$$\hat{w}_t = \text{argmax}_{w \in V}\, P(w|\mathbf{w}_{<t})$$

- The same output sequence for a given input sequence
- Therefore, no variability/diversity in output
- If the output is wrong, it will always be wrong (deterministic!)

44

Decoding by sampling (1/3)

Random sampling: Generate a word by 'sampling' from the distribution



Any word in the vocabulary
(based on the distribution as a prior)

$$\hat{w}_t = \cancel{\text{argmax}_{w \in V} P(w|\mathbf{w}_{<t})}$$

$$w_i \sim \mathrm{p}(w_i \mid w_{<i})$$

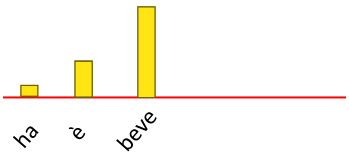
- Different output sequence for a given input sequence
- Some variability/diversity in output
- Because 'any' word can be generated, some outputs may be really weird

UNSW SYDNEY

45

Decoding by sampling (2/3)

Top-k sampling: Generate a word by 'sampling' from the top-k words



Any word from among the top-k words
(based on the distribution as a prior)

$$w_i \sim \cancel{\mathrm{p}(w_i \mid w_{<i})}$$

$w_i \sim \mathrm{p}(w_i \mid w_{<i})$.sort(reverse=True)[0:k]

- k is a parameter
- More controlled than random sampling; less deterministic than greedy decoding (when k>1)
- k=1 -> Greedy decoding

UNSW SYDNEY

46

Decoding by sampling (3/3)

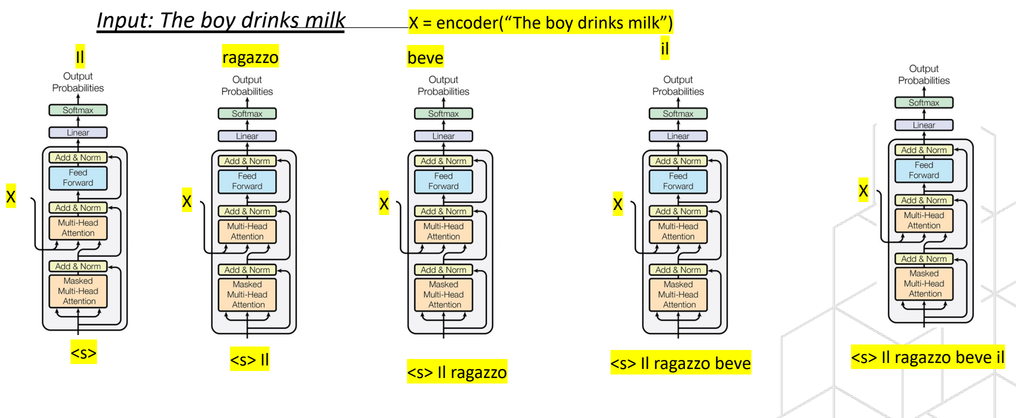
Other strategies

- top-p sampling: Probability mass p instead of number of words k
- Temperature sampling: Re-shape the distribution

47

Is this the only way to generate output?

Can we generate MULTIPLE possible sequences and then output the 'best'?
Recall Viterbi decoding!



48

Beam search

A useful algorithm for decoding using Transformer

Key idea:

Store multiple choices for incomplete sequences.

Pick the best sequence when you have generated the complete output sequences.

Typical beam widths for MT: 5-10

Fun fact: Beam search was first introduced in a PhD Thesis in 1976! (B T Lowerre, CMU)

UNSW SYDNEY

49

Pseudocode

beam width -> # options to consider

Start with an empty output string

frontier: 'leaf' on the generation path

Consider all frontiers and generate the next word

If you encounter end of sequence, decrement beam width

Select beam-width number of paths

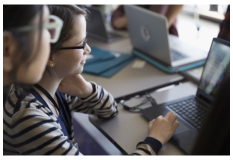
```
function BEAMDECODE(c, beam_width) returns best paths

  y0, h0 ← 0
  path ← ()
  complete_paths ← ()
  state ← (c, y0, h0, path)          ;initial state
  frontier ← ⟨state⟩        ;initial frontier

  while frontier contains incomplete paths and beamwidth > 0
     extended_frontier ← ⟨⟩
     for each state ∈ frontier do
        y ← DECODE(state)
        for each word i ∈ Vocabulary do
           successor ← NEWSTATE(state, i, y_i)
           extended_frontier ← ADDTOBEAM(successor, extended_frontier,
                                          beam_width)

     for each state in extended_frontier do
        if state is complete do
           complete_paths ← APPEND(complete_paths, state)
           extended_frontier ← REMOVE(extended_frontier, state)
           beam_width ← beam_width - 1
     frontier ← extended_frontier

   return completed_paths

function NEWSTATE(state, word, word_prob) returns new state

function ADDTOBEAM(state, frontier, width) returns updated frontier

  if LENGTH(frontier) < width then
     frontier ← INSERT(state, frontier)
  else if SCORE(state) > SCORE(WORSTOF(frontier))
     frontier ← REMOVE(WORSTOF(frontier))
     frontier ← INSERT(state, frontier)
  return frontier
```

Return multiple paths/sequences; not just one.

UNSW SYDNEY

50

Example



Demo time!

51

Part 4

LLM-based MT

Aharoni, Roee, Melvin Johnson, and Orhan Firat. "Massively Multilingual Neural Machine Translation." *NAACL 2019*.
Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research. 2020.
Wang, Yizhong, et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions." The 61st Annual Meeting Of The Association For Computational Linguistics. 2023.
Alves, Duarte M., et al. "Tower: An Open Multilingual Large Language Model for Translation-Related Tasks." arXiv preprint arXiv:2402.17733 (2024).

52

Multi-lingual Neural Machine Translation (MNMT)

Ability to translate from multiple source and target languages

Data:

- Parallel corpora for multiple pairs of languages
- Often English-centric

Typical techniques:

- Continued pre-training
- Instruction tuning

Examples of models:

- TowerLLM [1]
- Marian NMT [2]

[1] https://unbabel.com/announcing-tower-an-open-multilingual-llm-for-translation-related-tasks/
[2] https://marian-nmt.github.io/docs/

UNSW SYDNEY

53

.. so how does the model know which language to translate to?

Option 1: Special token at the beginning of the data

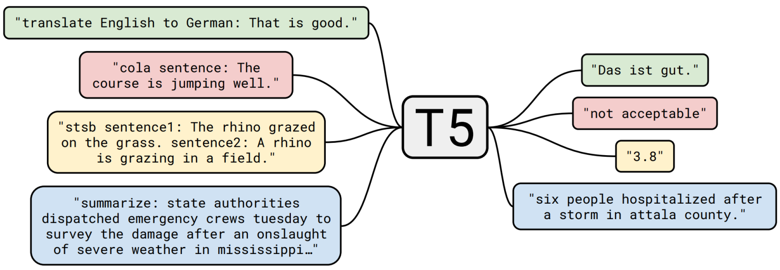
```
How are you? -> ¿Cómo estás?
<2es> How are you? -> ¿Cómo estás?
```

UNSW SYDNEY

54

Option 2: Instruction tuning

Convert all tasks to an “instruction: input -> output format”



55

Instruction tuning

Concatenate Instruction and Input to form the input to a decoder-only model

Fine-tune the model to learn to generate the correct output

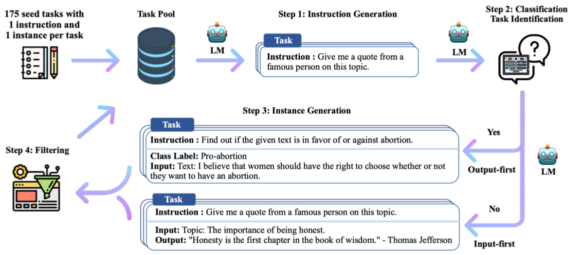


Figure 2: A high-level overview of SELF-INSTRUCT. The process starts with a small seed set of tasks as the task pool. Random tasks are sampled from the task pool, and used to prompt an off-the-shelf LM to generate both new instructions and corresponding instances, followed by filtering low-quality or similar generations, and then added back to the initial repository of tasks. The resulting data can be used for the instruction tuning of the language model itself later to follow instructions better. Tasks shown in the figure are generated by GPT3.

Classification tasks:
Generate the output label
Generate the instance based on the task and the output label.

Generation tasks:
Generate the input sequence
Generate the output based on the input sequence.

← What would this look like for MT?

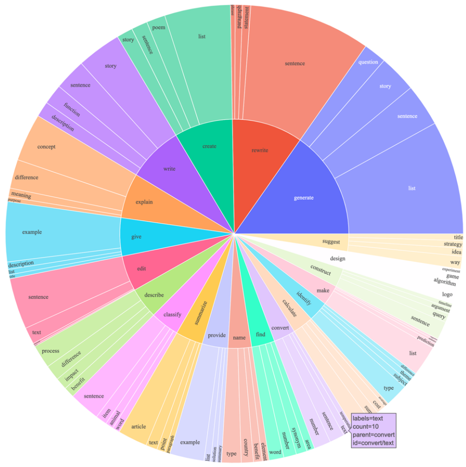
56

Stanford Alpaca: Instruction-following LLaMA model

Based on LLaMA: Large Language Model Meta AI

Instruction-tuned on a large set of tasks

Let's explore the github repo:
https://github.com/tatsu-lab/stanford_alpaca



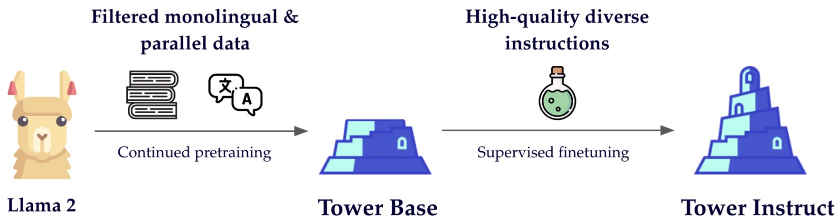
57

TowerLLM

Demo time!

Open Multilingual Large Language Model for Translation-Related Tasks

10 languages: English (en), German (de), French (fr), Dutch (nl), Italian (it), Spanish (es), Portuguese (pt), Korean (ko), Russian (ru), and Chinese (zh).



Do not pre-train from scratch.
Use available datasets to update the model.

https://huggingface.co/collections/Unbabel/tower-659eaedfe36e6dd29eb1805c

58

Part 5

Special cases of MT

Artetxe, Mikel, et al. "Unsupervised Neural Machine Translation." *International Conference on Learning Representations*. 2018.

59

Assumption so far

The assumption so far is that sufficient parallel data is available.
Not the case for many languages/dialects of languages.

"MT in low-resource setting"

Let's say: Arabic to Italian MT

60

Pivot-based MT

Case 1: Parallel data between source and target languages is not available

Arabic	English

English	Italian

Sentence in Arabic → Arabic -> English → English -> Italian → Sentence in Italian

Pivot language: English

The choice of pivot language may depend on language similarity.

61

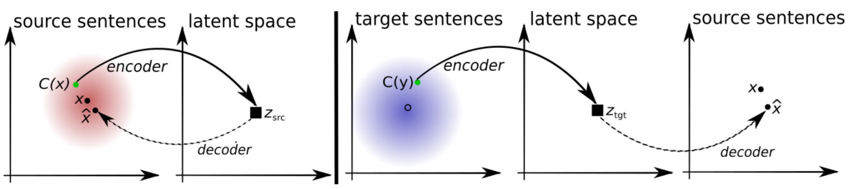
Unsupervised MT

Case 2: Parallel data between source and target languages is limited, but monolingual data in the two languages is available

What is the 'extreme' condition?

Arabic

Italian



C(x) -> Noisy input (Perturbations)

Learn a common embedding space for the two languages

Image from: https://arxiv.org/pdf/1711.00043.pdf

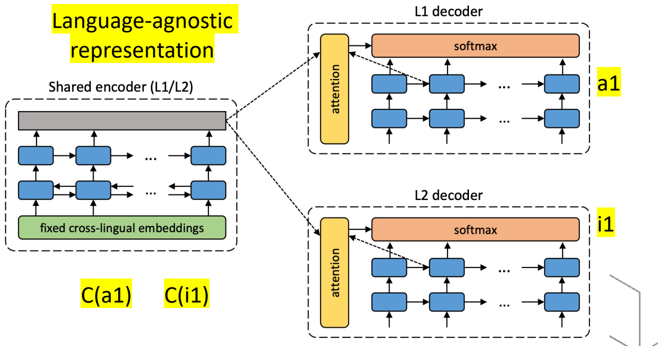
62

Unsupervised MT: Denoising & Backtranslation

Two alternating steps:
(1) Denoising
(2) Back-translation

Arabic
a1

Italian
i1



Note: The diagram uses pre-Transformer encoder/decoders. The two steps are thematically followed in many unsupervised NMT

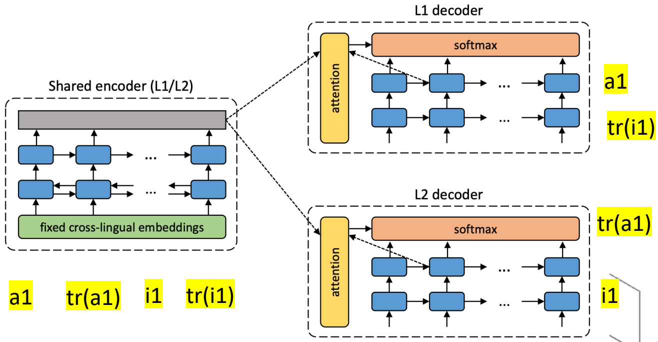
63

Unsupervised MT: Denoising & Backtranslation

Two alternating steps:
(1) Denoising
(2) Back-translation

Arabic
a1

Italian
i1



Try this library: https://github.com/artetxem/monoses

Note: The diagram uses pre-Transformer encoder/decoders. The two steps are thematically followed in many unsupervised NMT

64

Post-editing for MT

Post-editing: Humans may edit the output of an MT system

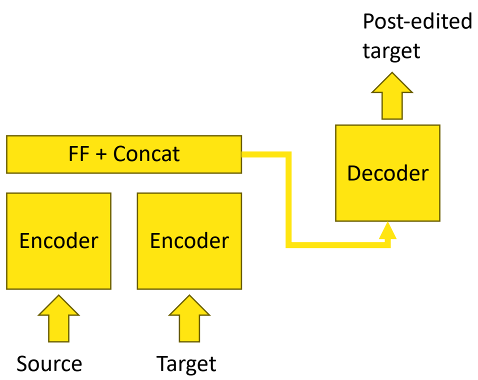
Can the post-editing be automated?

Can you think of rule-based methods?

65

Automatic Post-editing (APE) for MT

Automatic correction of the output of an MT system



66

Summary

Part	Key Concepts	Demos
Introduction	Terminology; history	Zero-shot MT
MT Evaluation	BLEU, METEOR, ROUGE, BLEURT	Libraries to compute the metrics
Statistical MT	Alignment & language models; IBM models	-
Transformer & MT	Transformer decoding: Greedy/sampling/beam search	OpenNMT tutorial Decoding strategies
LLM-based MT	Instruction tuning TowerLLM	Instruction tuning for MT
Special cases of MT	Pivot-based MT Unsupervised MT: Denoising & backtranslation Automatic post-editing	

67

Advanced/Optional Reading

Tay, Yi, et al. "Efficient transformers: A survey." *ACM Computing Surveys* 55.6 (2022): 1-28.

Maruf, Sameen, Fahimeh Saleh, and Gholamreza Haffari. "A survey on document-level neural machine translation: Methods and evaluation." *ACM Computing Surveys (CSUR)* 54.2 (2021): 1-36.

Ranathunga, Surangika, et al. "Neural machine translation for low-resource languages: A survey." *ACM Computing Surveys* 55.11 (2023): 1-37.

Apidianaki, Marianna. "From word types to tokens and back: A survey of approaches to word meaning representation and interpretation." Computational Linguistics 49.2 (2023): 465-523.

Tang, Yuqing, et al. "Multilingual translation with extensible multilingual pretraining and finetuning." arXiv preprint arXiv:2008.00401 (2020).

68

What about tasks where input and output are both sequences of significantly different lengths?

->(Week 9: Summarisation)

UNSW SYDNEY

69